Charles University in Prague

Faculty of Social Sciences

Institute of Economic Studies

MASTER THESIS

The Pharmacy-based Cost Group Model: Application in the Czech Health Care System

Author: **Bc. Tereza Hajíčková**

Supervisor: **PhDr. Jana Votápková**

Academic Year: **2014/2015**

Declaration of Authorship

The author hereby declares that she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, May 14, 2015

Signature

Acknowledgments

I wish to thank Jana Votápková, my thesis advisor, for her academic supervision, comments and suggestions that have profoundly influenced my way of thinking about the subject of this thesis.

Undertaking master studies has been a truly life-changing experience for me, and it would not have been possible to do without the support and guidance that I received from many people. Some special words of gratitude go to my friends who have always been a major source of support when things would get discouraging. Lastly, but most importantly, I wish to thank my parents, Jiřina and Jan, and my sister Marie for their unconditional love and support during my studies; my gratitude to them is beyond words.

Abstract

The risk adjustment model currently used does not adequately compensate insurers for predictable differences in individuals' health care expenditures in the Czech Republic. It then leads to financial inequality in the redistribution of funds to the insurance companies and causes their financial problems. This study introduces a PCG model as another method for risk adjustment and determines to what extent the predictive performance of the model can be improved when applied to Czech data. We analyze 10% of population sample in the Czech Republic in years 2011 and 2012. Our results confirm the appropriateness of the PCG model for the Czech environment. When the PCG variables are added to the demographic model, R^2 value of the prediction model increases from 2.03% to 13.87%.

JEL Classification	I13, I18,
Keywords	PCG model, health insurance, prediction of health care cost, risk adjustment

Author's e-mail	hajickova.tereza@gmail.com
Supervisor's e-mail	jana.votapkova@gmail.com

Bibliographic note

Hajíčková, Tereza. The Pharmacy-based Cost Group Model: Application in the Czech Health Care System. 58 p. Master thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Supervisor PhDr. Jana Votápková

Range of thesis 93 959 characters

Abstrakt

Současný model kompenzace rizika v České republice nedostatečně odráží variabilitu predikovatelných nákladů na zdravotní péči jednotlivých pojištěnců, čímž dochází k finančnímu znevýhodnění některých pojišťoven. Tato práce rozvíjí metodiku farmaceutických skupin (PCG model) a analyzuje přínos jejich zařazení do českého modelu, který je v současnosti používán. Pro analýzu jsou použita data 10% vzorku české populace v letech 2011 a 2012. Výsledky práce potvrzují vhodnost PCG modelu pro české prostředí. Zařazením PCG skupin do demografického modelu se R^2 hodnota modelu zvýšila z původních 2.03% na 13.87%.

Klasifikace JEL	I13, I18,
Klíčová slova	PCG model, zdravotní pojištění, predikce nákladů na zdravotní péči, riziková úprava

E-mail autora	`hajickova.tereza@gmail.com`
E-mail vedoucího práce	`jana.votapkova@gmail.com`

Contents

List of Tables

List of Figures

Acronyms

ATC	Anatomical Therapeutic Chemical
CDS	Chronic Disease Score
COPD	Chronic Obstructive Pulmonary Disease
DCG	Diagnostic Cost Group
DDD	Defined Daily Doses
GLM	Generalized Linear Models
GLS	Generalized Least Squares
HCC	Hierarchical Condition Category
ICD-9-CM	International Classification of Diseases, 9th Revision, Clinical Modification
MAPE	Mean Absolute Prediction Error
MARE	Mean Absolute Relative Error
OLS	Ordinary Least Squares
PCDS	Pediatric Chronic Disease Score
PCG	Pharmacy-based Cost Group
SUKL	State Institute for Drug Control
VZP	Všeobecná zdravotní pojišťovna
WHO	World Health Organisation
WLS	Weighted Least Squares

Master Thesis Proposal

Author	Bc. Tereza Hajíčková
Supervisor	PhDr. Jana Votápková
Proposed topic	The Pharmacy-based Cost Group Model: Application in the Czech Health Care System

Motivation In the Czech Republic, the public health care system is financed by insurance tax. The collected insurance funds are redistributed to the insurance companies according to the total number of enrollees. As risk factors, their gender and age are used. Significant risk of economic problems emerges for the insurance companies when their enrollees consume more health care than is predicted by the system.

There are other various approaches to evaluate and to predict medical costs in the health care system. Pharmacy-based Cost Groups (PCG) is one of the methods recently used. It is an outpatient morbidity measure based on information about chronic conditions deduced from the use of prescribed drugs.

We argue that PCG might be a feasible option for improving the risk adjustment system in the Czech Republic. Thus, the goal of this thesis is to examine whether the current Czech model can be improved by incorporating information on the presence of chronic conditions.

Hypothesis

1. Is the predictive performance of the PCG model for health care expenditures in the Czech Republic better than the performance of the demographic model? By how much?

2. Risk-adjustment by including the region of origin of individuals improves the predictive power of the model.

Methodology To apply the PCG model in the Czech environment, we will use data of drug usage by insured persons, their demographic characteristics and health care costs. To assign enrollees into PCG, we will basically use classification defined by Lamers and van Vliet (2004) which might be further adjusted to the Czech specifics of drug use.

Based on Lamers and Vliet (2003), firstly we will estimate the demographic model currently used. In this model, the independent variables are $age \times sex$ ($18 \times 2 = 36$ dummy variables) and the dependent variable is *health care expenditures*. Subsequently, we will estimate the PCG model. Therefore, the demographic model will be extended for dummy variables for *chronic conditions* based on previously defined classification. The model might be further modified with dummy variables for regions (NUTS 3-4) where individuals live.

We will calculate the proportion of explained variation for each regression model to assess their performance. Moreover, we will apply cost quintile analysis as risk groups may generate more precise predictions for different ranges of the cost distribution. (Fishman et al., 2003) Individuals will be grouped into equally populated segments based on their actual costs, and each quantile will be examining both statistically and graphically.

Expected Contribution A common challenge for health care reforms is to find adequate predictors of health care expenditures, in particular direct measures of health status among insured individuals. Different studies approved that PCG model has a good predictive power e.g. Lamers (1999a), Huber et al. (2013), Putnam et al. (2002). These results and implications may not be transferable to countries with diverse structure of population and health care system. Therefore, the contribution of our study is to apply the PCG model on Czech data to find out its possible application in the Czech health care system and its advantages over the currently used model.

Outline

1. Introduction
2. Environment overview
   (a) Czech demographic model
   (b) PCG model
3. Empirical analysis
   (a) Estimation of the demographic model
   (b) Estimation of the PCG model
   (c) Comparison of predictive performance of the models
4. Conclusion

Chapter 1

Introduction

All over the world, health care systems come under great pressure. Confronted with an increasing demand for health care, the health care expenditures continue to rise in many countries. As a result, policymakers are investigating methods to increase the quality and efficiency of care. There has been a growing belief that many inefficiencies stem from flawed provider payment systems creating perverse incentives for health care providers (Eijkenaar, 2013). In response, improved risk adjustment methods are introduced in the reimbursement for services.

Risk adjustment is a technique used to calculate the expected costs of treating a specific group. The model uses information about the members of that group to allocate funds among insurance companies that pay for health care of the people. An efficient risk adjustment scheme compensates insurers for predictable variation in individual health care expenditures between low-risk individuals (e.g. the healthy) and high-risk individuals (e.g. the chronically ill).

In case the risk adjustment scheme is not defined fairly, it introduces both efficiency problems as well as fairness problems. Under the former, the insurance companies are motivated to select low-risk individuals to save resources, under the latter insurance companies are overfunded by the government for low-risk groups and underpaid for high-risk groups. The goal is to ensure that an insurance company that attracts a sicker than average set of enrollees will receive greater resources and vice versa. Previous research has shown that past drug consumption is indicative of individual health condition and can be used as a health status proxy in risk adjustment models. Such models

have been reported to have higher predictive power than models bases solely on demographic factors such as age, gender and geographic location (Lamers and van Vliet, 2004; Fishman et al., 2003; Huber et al., 2013). Examples of pharmacy-based models are those used in the Netherlands or Slovakia.

This thesis introduces a pharmacy-based method for risk adjustment: Pharmacy-based Cost Group (PCG) model. PCG adjusters define chronic disease classes inferred from the prescription of drugs. The comparison of PCG and demographic model that is currently used in the Czech Republic is presented. We focus on the estimation of the prediction models, with the aim of explaining variation in individual health care expenses and to obtain accurate predictions, as far as is possible. We will further examine a use of regional variables in the PCG model. We argue that the demographic model currently used in the Czech Republic does not adequately predict the individual health care expenses. The PCG model and regional variables are expected to bring additional information about individual health status condition. Therefore, the underlying assumption is that using a PCG model and regional adjusters will substantially increase the predictive power of the model.

The sample used for the empirical analysis consists of 10% of the Czech population in 2011 and 2012. The proposed regression model includes demographic, regional and PCG variables and is estimated by Weighted Least Squares (WLS).

No profound analysis of PCG model and its application to the Czech system is available. As the results from previous studies may not be transferable to countries with a diverse structure of population and health care system, we will evaluate the appropriateness of the PCG model for the Czech environment. We will also analyze if the PCG model compensates health plans for the expected costs associated with the disease burden of their enrollees more accurately than the demographic model.

Thus, the thesis will answer the following questions:

1. Is PCG model valid prediction model of health care costs in the Czech Republic?

2. Does the predictive power of PCG model increase when regional adjusters are added?

3. Do PCG and regional models lead to a positive financial impact on insurance companies?

The remainder of this study is structured as follows. Chapter 2 describes risk in health care and presents the risk adjustment methods including the demographic and PCG models. Chapter 3 turns to describing the Czech health care system. Chapter 4 reviews the recent literature. In Chapter 5 the data description is presented, followed by methodology in Chapter 6. The results are summarized in Chapter 7 and are further discussed in Chapter 8. Finally, we conclude our findings and formulate points for further research in Chapter 9.

Chapter 2

Risk in Health Care

Policy makers across the world emphasize the importance of ensuring adequate health care for all. A significant barrier to this objective emerges as the cost of health care is frequently high relative to an individual's income and health care costs are not distributed evenly through the population. The health care cost falls disproportionately on disadvantaged populations, which are more likely to experience higher rates of unemployment and to have lower incomes that would not cover their health care needs. One group where high health risks emerge is the senior population. As the income of seniors is often low, it would not cover their total health expenditures.

Health care expenditures are also characterized by both large random variation as well as large unpredictable variation across individuals. At least 70-80% of the variation in annual health care expenditures among individuals is fundamentally unpredictable. Among the factors that influence the predictable part of the variation of health expenditures are age, sex, place of residence or past expenditures (Van Vliet, 1992; Smith and Witter, 2004).

Variations in health care expenditures create the potential for substantial efficiency gains due to risk reduction from insurance. It also raises important concerns about fairness across individuals with different expected needs for services (Culyer and Newhouse, 2000).

The solidarity principle, which is highly valued in Europe, implies that high-risk and low-income individuals receive an implicit subsidy to increase their access to health insurance coverage. In other words, both the high-risk and the low-risk

individuals pay the same premium. This mechanism called risk-pooling can be used to help spread the costs of health care over a population group (Roberts et al., 2008).

2.1 Risk pooling

The World Health Organisation (WHO) defines risk pooling as "the practice of bringing several risks together for insurance purposes in order to balance the consequences of the realization of each individual risk." (Reinhardt and Cheng, 2000)

Pooling ensures that the risk related to financing health costs is borne by all the members of the pool and not by each subscriber individually. Therefore, the main purpose is to share the financial risk associated with health interventions as an individual's need for health care is uncertain. Nobody knows if one will need costly health care in the future or if they will become poor. Uncertainty about getting sick and/or becoming poor in the future may motivate even the healthy and rich person to consider desirable some redistribution toward disadvantaged groups today, as a hedge against the risk of being in need of such transfers in the future.

The arguments supporting risk pooling in health care are primarily equity and efficiency. The equity arguments reflect the view that individuals should not face all the risk associated with their potential health care expenditure. The pattern of a burden of disease is related to poverty: the poor are the ones most in need for health care. Their low levels of income can result in an inability to seek treatment when it is paid, or there might be adverse consequences from seeking paid treatment (such as indebtedness). Thus, if the situation emerges, they are guaranteed adequate treatment and do not pay its full cost. The efficiency arguments emerge because pooling can help to improve population health, increase productivity, and reduce uncertainty associated with health care expenditure. With no pooling, poorer citizens might languish untreated and become a burden on society (Smith and Witter, 2004).

The access to health care should not be guaranteed only in time when the patient is in poor health, but also it should provide access to prevention. There are many positive externalities from increasing access to prevention and treatment.

Firstly, it limits the spread of infectious diseases. Secondly, considering the access to health care in wider economic terms, when a population is in good health, it leads to higher workforce productivity.

For the purpose of risk pooling and the process of collecting revenue, there is a health insurance market. The funds for health care are collected from the population (individuals and corporate entities) by a collection agent. The funding mechanism differs in various countries and often includes taxation, social insurance contribution, private insurance premiums or out-of-pocket payments[1].

After collecting the revenues, the funds are pooled i.e. all the prepaid revenues are accumulated together, and then are redistributed according to the defined allocation mechanism to the insurance funds. This contribution to insurance companies' budget associated with an insuree is called a capitation payment[2] or premium subsidies. As the health care expenditures needs of citizens vary considerably due to different individual health conditions, the capitation system therefore employs methods of adjustment. These methods seek to adjust per capita payments to reflect the relative expected health service expenditures for fund members on the basis of personal characteristics. As the allocation is being adjusted according to the risk profile of individuals, this process is referred to as risk adjustment.

2.2 Risk adjustment

It is important to define risk adjustment scheme well to effectively allocate the pooled funds to sickness funds. Thus, a system of risk-adjusted premium subsidies or risk equalization across risk groups should be implemented (Van de Ven et al., 2003) to ensure that each insurance fund has the correct relative level of allocated resources for the population which is responsible for (Mossialos et al., 2002).

The definitions of risk adjustment vary. Ellis (2008) defines risk adjustment as

[1] More information on the financing mechanisms may be found on http://ies.fsv.cuni.cz/sci/publication/show/id/3693/lang/en

[2] Since payments are calculated on a per capita basis.

the use of patient-level information to explain variation in health care spending, resource utilization, and health outcomes over a fixed interval of time, such as a year. Keenan et al. (2001) define risk adjustment as the adjustment of premiums paid to health plans (or to insurance companies) based on a formula employing individual level diagnostic and/or demographic information.

Risk adjustment in the health insurance market has developed from a concern to prevent cream-skimming. Barros (2003) defines cream-skimming as selection by providers (or entities responsible for health care provision) of those consumers expected to be profitable, given the system of risk-adjusted capitation payments. Because insurance funds are not allowed to refuse applicants, they have incentives to use the more subtle forms of selection, which may have negative effects. The insurance funds may provide poor service to the chronically ill and choose not to contract with providers who have the best reputations for treating chronic illnesses. In sum, selection may threaten good quality care for the chronically ill (Van de Ven et al., 2003).

Proper risk adjustment is an essential pre-condition for gaining the benefits of a competitive health insurance market. Risk-adjusted payments to health insurance companies should reflect each individual's expected cost of health services utilization. An ideal risk adjustment model levels the playing field, so that a insurance company is indifferent between accepting healthy or ill, young or old, rich or poor enrollees, because they are paid for the enrollees expected risk. Risk-adjusted capitation payments should therefore encourage sickness funds to concentrate more on cost-containment and efficiency instead of focusing on risk selection. The risk-adjusted capitation payments should explain predictable variations in annual per person health care expenditures, as far as these are related to health status. Among the risk adjusters commonly used are demographic and socioeconomic characteristics and health status proxies using the pharmaceutical or diagnostic information.

2.3 Risk adjustment methods

Among the risk adjusters commonly used are demographic and socioeconomic characteristics and health status proxies using the pharmaceutical or diagnostic information.

2.3.1 Demographic model

The demographic model uses as risk adjusters the individual age and sex categories. Dummy variables are used for the various age/sex categories. The independent variables used in demographic models are degree of urbanization, region, disability or employment status.

Various studies have shown that demographic variables are too crude as risk adjusters. The capitation system can be improved by extending the set of risk adjusters with measures that are more directly linked to health (Lamers and van Vliet, 2004).

2.3.2 Diagnosis-based models

The essence of diagnosis-based models is to classify individuals into cost groups, based on a limited set of diagnoses from hospital admissions in the preceding year.

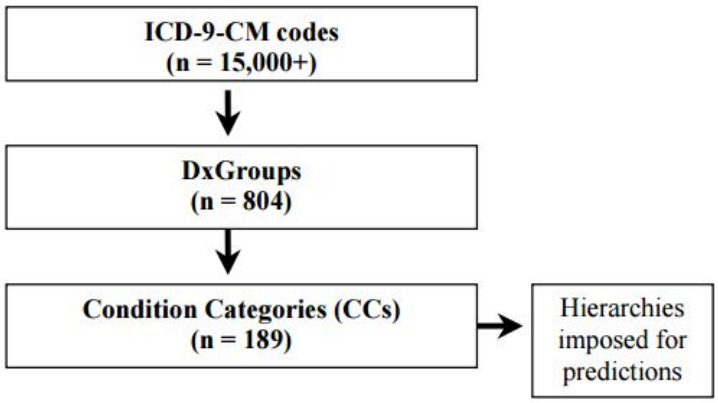
The Diagnostic Cost Groups (DCGs) were developed in the United States by Ash et al. (1990). They developed DCG using empirically determined similarities in the future costs of individuals hospitalized for different reasons. They aggregated codes of International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) into nine diagnostic subgroups. The DCG model improved R^2 value to 4.5% compared to the value of 0.5% for the demographic model. This model was further enhanced as described in Ellis et al. (1988); Ash (1997).

Pope et al. (2000) set a new DCG-HCC clinical classification system. This DCG-HCC system classifies each of the ICD-9-CM codes into diagnostic grouping called DxGroups. These groups are further aggregated into Hierarchical Condition Category (HCC) based on major diseases that are similar both clinically and costly. This mechanism of aggregation is presented in Figure 2.1.

Pope et al. (2000) reported the R^2 value of 11.2% for the DCG-HCC models compared to the 1% when only the demographic variables are used.

In Europe, the DCG methodology was applied in Lamers (1999b) which compared the demographic model, DCG model based on one year of observation and also DCG model based on three years of observation. Demographic model reported

Figure 2.1: DCG aggregation of ICD-9-CM codes



Source: Pope et al. (2000)

R^2 value of 3.78%, DCG model of one year 6.48% and DCG model based on three years 8%. Prinsze and van Vliet (2007) combined DCG and PCG in a single prediction model which again improved the R^2 value to 22.8%.

2.3.3 Pharmacy-based models

A weakness of the demographic model is that predicted expenses are not adjusted for the large differences in individual health within each age/sex group. A way to improve the predictions is to extend the set of risk adjusters with health status proxy (Van de Ven et al., 2004).

Pharmacy-based models assume that the prescription data well captures the morbidity conditions. The type of medication reveals the severity of the condition being treated, and the filled prescriptions also represent the conditions that were judged by the patient to be serious enough to seek treatment. These conditions suggest that these chronically ill patients will require ongoing treatment and subsequently higher expenditures. Therefore, the prescription data is considered to be a reliable instrument for predicting future health care utilization (Kuo et al., 2011).

Several pharmacy-based morbidity measures have been developed and applied for predicting health care utilization. The Chronic Disease Score (CDS) is the very earliest one and was designed by Von Korff et al. (1992) in the United States. It used data only for an adult population. The authors established a set of dummy variables that indicate a pharmacy prescription during a base period

for a medication representing particular chronic diseases. In total 28 different conditions were distinguished. This score was evaluated in terms of its stability over time and its association with other health status measures. It predicted hospitalization and mortality in the following year after controlling for age and gender and moreover CDS showed a high year to year stability. This CDS model containing 28 binary variables together with age and sex explained 10% of the variation in total health expenditures of adults enrolled in a health maintenance organization in the next 6-month period. Age and sex alone (demographic model) explained 3% of the variation in total charges.

Fishman and Shay (1999) focused on the evaluation of a pediatric risk assessment model for children population in the period 1992-93 and developed Pediatric Chronic Disease Score (PCDS). They employed a methodology similar to that used in the development of the CDS. Drugs commonly used in ambulatory settings for the treatment of specific chronic conditions were grouped into 26 diagnostic categories. The model including PCDS explained 17% of the prospective individual variation in costs and significantly outperformed the demographic model with R^2 equal to 1.4%.

Fishman et al. (2003) revised and expanded the CDS and PCDS. The original established CDS was not intended as a capitation payment adjuster. Hence, it included several categories that authors found inappropriate in a model used for financial purposes as some drugs are prescribed less systematically e.g. pain drugs. The new model, called RxRisk, was an all-ages and market segment pharmacy-based risk assessment model. Fishman et al. (2003) used total costs as a proxy for medical risk. They assumed that medical risk during the examined time period is a function of each individual's age and sex, the source and extent of their health insurance (commercial, Medicare, or Medicaid) and the set of chronic conditions they are being treated for. These chronic conditions were measured by the RxRisk algorithm during a previous period, an updated CDS. RxRisk had an R^2 of more than 9% for the estimation sample. The demographic instrument explained relatively little of the prospective variance in cost with an R^2 of 3.9%.

These pharmacy-based risk adjustment tools have been tested and were found to be valid in predicting future health care utilization in the US, but most of them incorporate a coding algorithm that is specifically designed for the United States. Kuo et al. (2011) verified that the approaches incorporating a coding

algorithm that requires the medication data to be coded using the US National Drug Codes or the American Hospital Formulary Service drug codes can be applied also with Anatomical Therapeutic Chemical (ATC) algorithm and are able to explain the variations in health care utilization. This study also revisited the CDS by adding new medications to the original morbidity classes, and by included new disease categories which may be more appropriate to capture morbid conditions for different ethnic populations. The performance of the pharmacy-based metric model was tested using a sample representing the entire population of Taiwan. The resulting R^2 in the demographic model was only 3.5% whereas in case of model including pharmacy-based metric was 15.1%. By adding Elixhauser's comorbidity index[3] to the model, R^2 rose to 19.2%. The study also compared the pharmacy based metric with diagnosis-based models and found out that pharmacy based models explain or predict the health care costs better.

PCG model

The PCG model works on the assumptions of the CDS that drug consumption is a valid predictor of health care expenditures. The PCG model further adjusts the CDS for the European health specification. In Europe, the ATC coding is used. Therefore, prescription drug items are coded based on the WHO ATC classification system (Huber et al., 2013).

The PCG is an outpatient morbidity measure based on prior use of prescribed drugs. According to the PCG model, ATC codes are assigned to different chronic diseases. The essence of the risk adjustment using PCG is that enrollees are classified into clinically homogeneous groups based on the prior use of pharmaceuticals. Using the ATC classification system, the active substances are divided into different groups according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

Lamers (1999a) adjusted the US classification of medications representing chronic conditions that underlies the PCG model to the Dutch situation. He

[3] A comorbidity index based on an approach to identifying comorbidities and separates them from the primary reason for hospitalization. It results in an expanded set of comorbidities. It should be used carefully as an index because comorbidities affect outcomes differently among different patient groups (Elixhauser et al., 1998).

extended the study on the CDS using automated outpatient pharmacy data of one Dutch sickness fund. The 28 original chronic conditions were clustered into seven PCGs according to empirically determined similarities in future costs. The clustering of conditions into these seven PCGs almost did not affected the predictive power of the model. The PCG model explained 10% of the differences in next year's expenditures between individuals, which was almost twice the R^2 of a model containing only demographic variables.

Lamers and van Vliet (2004) newly derived the classification of drugs (see Table 2.1) and used it to estimate a capitation model containing demographic variables and information on chronic conditions. Based on these ATC-codes, persons with claims for medications were assigned to 20 chronic conditions. A demographic model alone explained 5.0% of the differences in health care expenditures in 1998 among individuals. The R^2 value of the chronic conditions model was 9.2%.

Table 2.1: Pharmacy-based cost groups

Chronic condition	**ATC-code**	**Description of ATC-code**
Epilepsy	N03A (excluding N03AE01)	Antiepileptics
Hypertension - high	C02	Antihypertensives: antiadrenergic agents, ganglion-blocking, peripherally acting, other antihypertensives
	C08	Calcium channel blockers
	C09A, C09B	Angiotensin-converting enzyme (ACE) inhibitors
Hypertension - low	C03A, C03EA01	Low-ceiling diuretics, thiaziden
	C07	Beta-blocking agents
HIV/AIDS	J05AB06, J05AD01, J05AE, J05AF	Ganciclovir, foscarnet, reverse transcriptase inhibitors
Tuberculosis	J04A	Drugs for treatment of tuberculosis
Rheumatologic conditions	M01CB, M01CC01, P01BA02,	Gold preparations, penicillamine, hydroxychloroquine,
	L01BA01, A07EC01	methotrexate, sulfasalazine
Hyperlipidemia	C10A	Cholesterol and triglyceride reducers
Malignancies	L01 (excluding L01BA01),L03AA02/03/10, A04AA	Antineoplastic agents, filgrastim, molgramostim, lenograstim, serotonin (5HT3) antagonists
Parkinson's disease	N04B	Dopaminergic agents
Renal disease	B03XA01, V03AE01	Erythropoietin, polystyrene suplhonate

Continued on next page

Table 2.1: Pharmacy-based cost groups (Continued)

Chronic condition	**ATC-code**	**Description of ATC-code**
Cardiac disease/ASCVD/CHF	C01	Cardiac therapy: Cardiac stimulants and glycosides, Antirrhythmics: class I and III, vasodilators used in cardiac diseases
	C03C, C03EB01	High-ceiling diuretics
Glaucoma	S01E	Antiglaucoma preparations
Peptic acid disease	A02A, A02B	Antacids. drugs for treatment of peptic ulcer
Cystic fibrosis	A09AA02	Multienzymes
Transplantations	L04AA01/5/06, L04AX01	Ciclosporin, tacrolimus, mycophenolic acid, azathioprine
Respiratory illness, asthma	R03	Antiasthmatics
Thyroid disorders	H03A, H03B	Thyroid preparations, Antithyroid preparations
Gout	M04A	Antigout preparations
Crohn's and ulcerative colitis	A07EC (excluding A07EC01)	Mesalazine, olsalazine
Depression	N06AB, N06AE	Selective serotonin reuptake inhibitors
	N06AF, N06AG	Monoamine oxidase inhibitors
	N06AX	Other antidepressants
Diabetes I	A10A	Insulins
Diabetes II	A10B	Oral blood glucose lowering drugs

The PCG model should identify only chronically ill people and should not include the prescription addressed to a minor temporary health problem. The studies of Lamers (1999a); Lamers and van Vliet (2004) classified a person to a PCG based at least on four prescriptions per year. Lamers and Vliet (2003) used as a threshold prescribed sum of Defined Daily Doses (DDD) comparing the thresholds of 91 DDD and 181 DDD. Using DDD eliminates the possible form of manipulation in this context e.g. a patient could receive four times a prescription for 1 week instead of a prescription for a whole month.

Huber et al. (2013) updated the model to the Swiss environment and created a model including 22 chronic conditions, age, sex, language area, managed care, deductible and accident coverage as independent variables to predict to total health care cost. They used two-part regression models. In the first part, logistic regression si used to determine the probability of incurring health care costs per patient/year. In the second part, Generalized Least Squares (GLS) with gamma error distribution and linear link function were used to estimate annual health care expenditures. They calculated the expected total health care costs

by multiplying predicted values from the first and second stage. The model explained 17.9% for the people aged 18-65years and 14.1% for persons aged 65 years and up. The model again outperformed the demographic model which reached R^2 of 2.5% and 6%, respectively.

Table 2.2 summarizes the application of different pharmacy-based models in the literature.

Table 2.2: Application of Pharmacy-based models in the literature

Country	**Author**	**Name of model**	**Sample**	**Variables**	**Method**	**R2**
US	Von Korff et al. (1992)	CDS	adult	28 chronic groups, age, sex	OLS	10%
US	Fishman and Shay (1999)	PCDS	children	26 chronic groups, age, sex	OLS	17%
US	Fishman et al. (2003)	RxRisk	all-ages	57 chronic groups, the source of health insurance	WLS	9%
NL	Lamers (1999a)	PCG	all-ages	7 chronic groups, age, sex, degree of urbanisation, disability status	OLS	10%
NL	Lamers and van Vliet (2004)	PCG	all-ages	25 chronic groups, age, sex, degree of urbanisation, disability status	OLS	9.2%
TW	Kuo et al. (2011)	PCG	all-ages	32 chronic groups	log-transformed OLS	15.1%
CH	Huber et al. (2013)	CDS	18-65 yr	22 chronic groups, age, sex, language area, managed care, deductible,	logistic regression +GLS	17.9%
			65 yr up	accident coverage		14.1%

Our analysis will compare the PCG model with the demographic model currently used in the Czech Republic. The reason we have decided to explore the PCG model carefully is that it yields significant improvements in the prediction accuracy. Moreover, it is easy to implement as the data of drug consumption are available in the Ministry of Health.

2.4 Risk adjustment in other European countries

Risk adjustment is applied differently in various health care systems. In the next sections, we describe several advanced systems of risk adjustment used in the present.

2.4.1 The Netherlands

The Netherlands is the pioneer in risk adjustment in Europe. It serves as a model for other countries (Van Kleef et al., 2013).

The Netherlands has implemented the risk equalization model since 1993. Within the current scheme, a prospective payment is made to insurers for each enrollee on their list, depending on risk characteristics of that enrollee. The following risk characteristics were added to the model: age interacted with gender (1993), region (1995), source of income interacted with age (1999), PCGs (2002), diagnoses-based cost groups (2004), socioeconomic status interacted with age (2008) and prior multiple-year high costs (2012). Overall there was 126 risk classes in 2012.

The current model comprises 40 classes for age and gender (20 classes for men and 20 classes for women). The age classes are 0, 1-4 and then 5-year groups up to the age of 90 and finally a class for people of 90 years and older. In addition, the model includes 10 clusters of regions. This urbanization criterion divides about 4000 Dutch postal codes into ten different groups, and this allocates every Dutch person uniquely to one of the ten groups.

The model also includes 17 classes for source of income, interacted with age. The following four sources of income are distinguished: self-employment, disability benefits, social security benefits and other (including employment). Each of these groups is interacted with age groups 18-34, 34-44, 45-54 and 55-64 years. Enrollees in the age groups 0-17 and 65 years or more are classified in one separate class.

The first direct proxy for health status used 26 PCGs. Enrollees are categorized in one or more of the PCG if they received at least 180 DDD of a certain

pharmaceutical in the preceding year. People who are not classified to any PCGs are assign to a group "no PCG". An individual can be classified to more than one PCG. As another direct proxy for health status, the model includes 14 DCGs. Using DCGs, individuals are classified into cost groups based on inpatient diagnoses from the previous year. Insurees with multiple diagnoses are classified in only one DCG (that with the highest costs). In 2013, the Dutch government made the DCGs also dependent on diagnostic information from prior outpatient visits in hospitals (not only outpatient). Each enrollee can be classified into one DCG at the maximum.

Furthermore, the model comprises 12 classes for socioeconomic status, interacted with age. This classification is based on income, a number of household members and age. For each enrollee, the income level is calculated as the household income divided by the number of household members. There are three groups: the bottom 30% of the income distribution, the middle 30-70% and the top 30%. Enrollees living in a household with more than 15 members are classified into a class independent of their income (the assumption being that they are living in a nursing home, an institution for handicapped or similar facility. Each of these socioeconomic classes interacts with three age groups (0-17, 18-64 and 65 years and more).

Since 2012, the model includes separate classes for enrollees with high costs in three previous years as those with high cost probably suffer from a chronic disease. Enrollees are classified into six cost groups if their health care expenses in year t-1, t-2 and t-3 are in top 1.5%, 4%, 7%, 10% or 15%.

The value of R^2 reaches 29.6% for the Dutch model (Van Kleef et al., 2013).

2.4.2 Germany

Originally, risk adjustment between the insurance funds in Germany was based on socio-demographic factors: age, sex, invalidity pension status and type of entitlement for sickness allowances.

There are single-year age groups up to 90 years for calculating standardized expenditures separately for men and women. The recipients of invalidity pension form another risk class. The entitlement for sickness allowances creates other

three groups: no entitlement, entitlement after 6 weeks and entitlement from the first day of absence from work.

In 2009, the new risk adjustment scheme was implemented called DCG-HCC model. This model separates DCGs into disease hierarchies that allow classifying individuals with multiple (unrelated) diagnoses into multiple risk classes[4]. There are 50-80 diseases selected which has to exceed 1.5 times the average per capita expenditure of all insured (cost threshold). There are 366 diseases based on the 781 DCGs of the HCC classification model. Furthermore, there are two criteria to be fulfilled in order to distinguish only severe and chronic diagnoses:

1. a disease has to be diagnosed in at least two different quarters of the year of observation in at least 50% of cases
2. at least 10% of the cases had to be hospitalized

After applying these criteria, the 80 most expensive diseases are selected and integrated into the risk adjustment model.

The R^2 value of the model was 23.9% in 2011 (Buchner et al., 2013).

2.4.3 Slovakia

Slovakia is a country that has recently introduced a PCG model into practice. Initially, the risk adjustment system distinguished only the economical active and inactive insured persons between 1995 and 2004. With the reform in 2004, the age and gender were introduced as risk adjusters. The age cohorts were divided by five years up to the age of 80 and again interacted with sex. In 2012, the economic activity status was added, and the PCG system was introduced. The Slovakian system used 24 PCGs in 2014. The threshold number of DDD is 181, the same threshold as used in the Netherlands. The Slovakian system does not allow comorbidity. In case that a person is classified into more than one PCG, only the most expensive group is taken into account.

Adding PCGs to the risk adjustment model raised the predictive power of the model from 3.19% to 19.58% (Szalayova, 2012).

[4] It is different to the Dutch model where diagnoses are grouped in DCGs only by similarities in follow-up costs and individuals can be classified in one DCG only.

Chapter 3

Czech Health Care System

The Czech Republic has a system of social health insurance based on compulsory membership in one of the health insurance funds. It is characterized by high level of solidarity which means that the health care provided is at the same level for everybody regardless how much he contributes.

The social health insurance contributions are wage-based payments paid by employers, employees, self-employed individuals and for groups of inactive people (unemployed, students, retired people, ...) paid by the state.

3.1 History of the Czech health care system

Czech (and, at that time, also Slovak) health care was characterized by universal coverage, national tax financing along with national ownership and control of health providers. The new system of compulsory social health insurance was introduced in 1992. The insurance was initially managed by a single insurer the General Health Insurance Company - Všeobecná zdravotní pojišťovna (VZP), which was responsible for collection and pooling of premium, as well as purchasing health care services for the entire population. Parliament approved another law enabling the foundation of competing non-profit making insurers established as public institutions in 1992. The first started operating in 1993. They were primarily organized around large employers or by industry sectors and were thus called branch or employers' health insurers. Their number reached 27 in 1995 and then decreased rapidly as many of them experienced financial

problems (Kutzin et al., 2010). By the year 2012, the number of insurers stabilized at seven (Table 3.1), but 59% of the population still belongs to the VZP.

Table 3.1: Czech insurance companies

ID	Insurance company
111	Všeobecná zdravotní pojišťovna
201	Vojenská zdravotní pojišťovna
205	Česká průmyslová zdravotní pojišťovna
207	Oborová zdravotní pojišť. zam. bank, poj. a staveb.
209	Zaměstnanecká pojišťovna Škoda
211	Zdravotní pojišťvna ministerstva vnitra ČR
213	Revírní bratrská pokladna, zdrav. pojišťovna
217	Metal-aliance

The subscribers to the branch insurers were primarily employed citizens and relatively younger and healthier individuals, retired and non-active citizens stayed subscribed to VZP. As each insurer collected premium (set as a payroll tax) independently, the difference in the risk composition of membership caused a rapid deterioration of the financial situation of the VZP. Therefore, the maintenance of isolated pools became unsustainable.

In 1994, a national pooling arrangement was introduced through a simple risk-adjustment mechanism administered by the VZP. Approximately 70% of collected funds - 60% of collected premiums and the whole payment from the state budget on behalf of non-working people (state insurees) - were subject to redistribution between insurers. The funds were redistributed between insurers according to the number of state insurees enrolled with each of the insurers with adjustment for age. Within the state insurees, two age categories were recognized - below and above 60 years and were assigned different weights in the risk adjustment formula. The group above 60 years was used with a triple weight than the younger group.

This arrangement enabled a more equitable division of available resources between the VZP and other health insurers, but it did not eliminate incentives for cream-skimming. Enormous differences (up to 50%) existed in average premiums collected from the economically active population, thus presenting an important disadvantage for insurers with higher shares of lower-income insurees. The age structure of the VZP's clients, combined with the low level

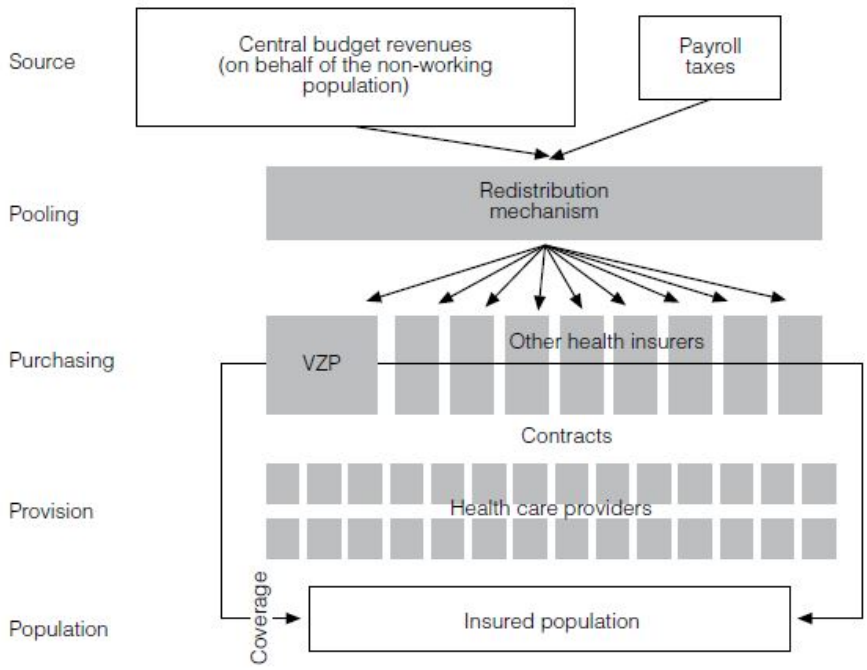
of state premium payments, contributed to its repeated deficits. Conversely, the other insurers reported mostly positive or at least neutral results.

From 2004 to mid-2006, a new risk-adjustment process was gradually implemented. Since then, all collected funds are subject to redistribution that combines a more refined ex-ante formula and an ex-post partial compensation of expensive cases.

3.2 Current situation in the Czech Republic

The reform of 2004 introduced complete pooling of the state payment and all collected premiums, which are redistributed among insurers on a capitation basis. The pooling mechanism is captured in the Figure 3.1

Figure 3.1: Pooling mechanism of funds



Source: Kutzin et al. (2010)

The new capitation formula is based on age (grouped into 5-year categories) and on gender, forming a total of 36 groups. Each year, Ministry of Finance and Ministry of Health issue new regulation with cost indices (weights) for each age/gender group which determine the redistribution of funds to each insurance funds depending on their composition of policyholders.

Each insurer reports the total amount of its collected premiums, as well as the number and age structure of its insured individuals on a monthly basis. State payments for economically non-active citizens flow directly to a special account operated by the VZP under the supervision of other insurers and the Ministry of Health and Finance. The account's manager then calculates the total amount of income (collected premiums + state payment) per "standardized" insured individual (woman aged 15-20 years). The amount of money allocated to each insurer then reflects the age/sex structure of the insuree.

The cost indices for the year 2015 are summarized in the Table 3.2.

Table 3.2: Cost indices

Age group		**Cost weights**	
From	To	Man	Woman
0	5	1.4572	1.3104
5	10	0.8549	0.7337
10	15	0.8500	0.9178
15	20	0.7680	1
20	25	0.6433	0.9771
25	30	0.7293	1.2726
30	35	0.8143	1.4144
35	40	0.9038	1.3463
40	45	1.0329	1.3542
45	50	1.2482	1.5727
50	55	1.5980	1.8583
55	60	2.1939	2.1331
60	65	2.8450	2.4374
65	70	3.6289	3.0240
70	75	4.2934	3.5835
75	80	4.8464	4.1361
80	85	4.9879	4.5522
85	and more	5.1021	4.9410

Source: 2012, Directive n. 442/2012 Sb.

Cost indices are determined as a share of average costs of an insured individual in the age/sex group when the costs of expensive clients[5] are subtracted. The average costs in each group are set as an average costs of all insurance funds for an insuree in the group. The average costs of a standardized insuree

[5] Their costs are 25 times higher than the average cost.

are determined as an average of costs of all insurance funds for a women aged 15-20 years.

In addition, the system includes an ex-post partial compensation of expensive clients. If the annual costs of an insured individual are 25 times greater than the average annual costs per client in the entire social health insurance system, the insurer is compensated for 80% of the costs over the limit once a year i.e. when the previous year's financial results are published. This compensation is intended to protect the health insurance funds from unexpected fluctuations in expenditure.

Chapter 4

Literature Review

The efficiency of different risk adjustment models has been widely assessed in the literature. The linear regression model using the Ordinary Least Squares (OLS) and WLS estimators has been commonly used (Fishman et al., 2003; Lamers and van Vliet, 2004; Zhao et al., 2005; Van de Ven et al., 2014). Applying OLS models on untransformed data for predicting individual expenses has been however discussed widely (Mihaylova et al., 2011; Lin, 2008; Jones, 2010), because OLS may not fit the distributional properties of health care expenses very well.

Jones (2010) and Veazie et al. (2003) assert that medical expenditures typically feature a spike at zero (i.e. nonusers) and a strongly skewed distribution with a heavy right-hand tail. This non-normality stems from the fact that, due to clinical complications and comorbidities, only a small minority of patients attract substantial and costly services. Their individual treatment is very expensive, creating outliers in the right-hand tail of the distribution. In econometric models of health care costs, the error term will be heteroskedastic reflecting the heterogeneity across patients. Due to these characteristics, reliance on a linear model estimated by OLS is inappropriate.

The presence of a substantial proportion of zeros in the data and a right-skewed distribution for users has typically been addressed by using a two-part model. This method distinguishes between a binary indicator, used to model the probability of any costs, and a conditional regression model for the positive costs (Huber et al., 2013; Powers et al., 2005).

Alternatively, the transformation of the dependent variable is used. The common

transformation included the log-transformation, but the square-root transformation and other power functions are applied as well (Jones, 2010). Using a transformation of cost data typically reduces skewness and, therefore, makes the distribution more symmetric and closer to normality. Parameter estimates from the regression of the transformed dependent variable are however on a different scale while analyst typically want results expressed in terms of actual costs. Re-transformation to the original scale is problematic due to re-transformation bias. Dunn et al. (2003) claim that a smearing factor is therefore necessary to apply which is problematic in cases involving heteroskedasticity in the data on the transformed scale. Note that Ellis et al. (2013) criticized the log-transformation for severe overestimation of the upper expenditure tail for the re-transformed data (raw). They justify the use of OLS explaining that the mean prediction absolute error from the log model was twice that from OLS.

Manning and Mullahy (2001) applied the Generalized Linear Models (GLM) for the health care cost modeling, which accommodates skewness through variance weighting. The advantage of GLM approach is that it does not require any retransformation (as needed in log-transformed OLS), the dependent variable is an expected value and is modeled on its original scale (Lin, 2008). According to Ellis et al. (2013) however GLM substantially reduces the weight that is put on observations with very high expenditures.

Powers et al. (2005) performed a comparative study of currently used models (OLS, log-OLS, logistic/OLS two-part model, logistic/log-OLS two-part model, logistic/GLM two-part model) and demonstrated that simple OLS model perform equivalently, in some cases with superiority to the variety of advanced econometric models presented in his study. The R^2 values for the log-OLS model, logistic/OLS, logistic/GLM two-part model was lower compared to the OLS method and when compared with the logistic/log-OLS two-part model the R^2 values were similar.

Supporting Powers et al. (2005), Fishman and Shay (1999) concludes that OLS regression performs well as a forecasting model relative to more sophisticated functional forms as it always yields unbiased estimates of parameter means. Wooldridge (2012) explains that heteroscedasticity does not cause bias or inconsistency in the OLS estimators, and the R^2 is also unaffected by the presence of heteroscedasticity. On the other hand, the estimates of the variance

are biased. Since the OLS standard errors are based directly on their variances, they are no longer valid for constructing the confidence intervals and t-statistics which then needs to be considered during the analysis. When a large sample is considered, non-normality is not a problem (Greene, 1997).

Health systems are country-specific. The results and implication of the previous studies may not be directly transferable to the Czech health care system. Unfortunately, no sophisticated analysis of risk adjustment mechanisms is available in the Czech environment. Thus, our study will contribute to the existing literature by comparing two methods of risk adjustment i.e. demographic model and PCG model, using data from the Czech health care system.

Chapter 5

Data

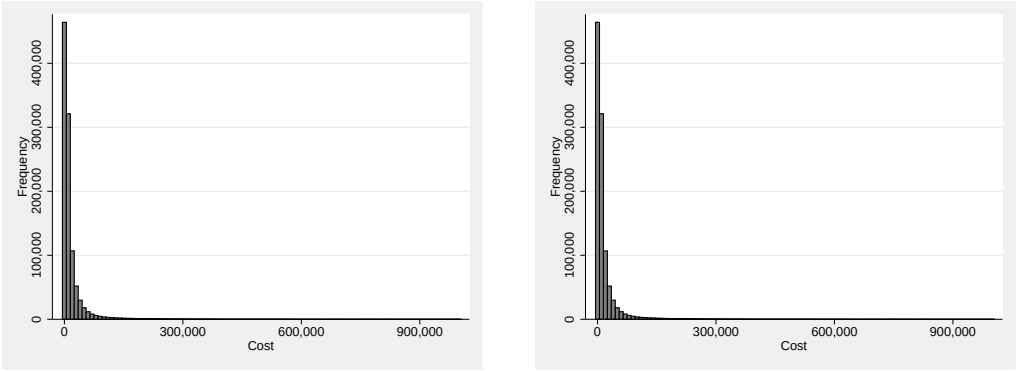
This chapter describes the data set, variables used and provides descriptive statistics.

5.1 Data description

The empirical analysis is based on a sample of 1 058 197 observations in the period 2011-2012 which represents 10% of all enrollees in the Czech health insurance program provided by the Czech Ministry of Health. Heath characteristics included annual health care expenditures in CZK and information about the number and type of prescribed drugs. Each drug had a specific ID, which is assigned by the State Institute for Drug Control (SUKL). The Ministry also provided us with the information on ATC specification and its DDD for each drug. Demographic characteristics included sex, date of birth, region and also the insurance company of the individuals.

5.1.1 Health care cost

The dependent variable in each of the estimated models is individual annual health care expenditure. As in Lamers and van Vliet (2004), for each partial-year enrollee who dis-enrolled or died during the given period, costs were raised to annual rates and eligibility fractions were used as weights. For example, an individual who died in June of year t and expended 10 000 CZK costs was



(a) Cost distribution in 2011 (b) Cost distribution in 2012

Figure 5.1: Cost distribution

assigned weight 0.5 and 20 000 annual expenses. By applying this procedure, mean predicted expenses in year t equals mean observed expenses in year t.

Summary statistics of individual health care expenditures are presented in Table 5.1.

Table 5.1: Summary statistics of health care costs

Year	**Mean**	**Std. dev.**	**Min**	**Max**
2011	19 811	72 298	0	14 925 653
2012	20 285	75 123	0	15 157 077

The distribution of the costs up to 1 000 000 CZK in 2011 and 2012 is presented in the Figure 5.1.

The costs are right-hand truncated at 1 000 000 in the Figure 5.1. Nevertheless, the distribution still has a marked long right tail. Even if we lower the threshold for truncation to 100 000 CZK, we find that only 3.16% of the sample exceeds this threshold. This truncation was done only for the purpose to display better the distribution (with larger-scale), but the further analysis in this study is done on the whole sample to avoid the bias in the estimates.

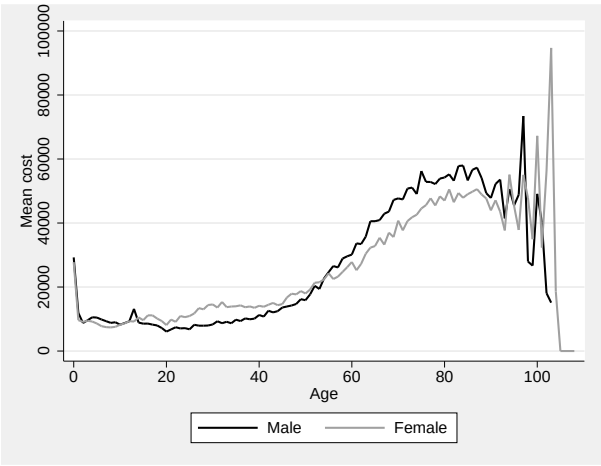
There are some extremely expensive cases in the sample that might be considered as outliers; nevertheless we keep them in the sample. Although removing outliers observation would likely improve the overall fit of the models, it is not the primary objective of our analysis. We assess the ability of the model to predict the future total costs for high-cost health care participants that often contribute

a disproportionate amount of cost to the system. This can by done only by keeping the outliers in the estimation sample. (Powers et al., 2005)

5.1.2 Age and gender

In the current risk adjustment model used in the Czech Republic, 36 groups are included to adjust for cost differences among age and sex groups. The population is separated by sex and divided into 5-year intervals with two exceptions. Firstly, the oldest age group comprises people aged 85 years and more consistent with the current practice in the Czech Republic. Secondly, we decided to divide the youngest group (0-5 years) into two categories: up to 1 year and 1 to 5 years. The reason is that there is a significant difference between the average cost of a newborn and the older children. This difference, as well as the differences between costs of woman and men, are depicted in the Figure 5.2.

Figure 5.2: Mean age cost



Each group is included in the regression as a dummy variable with the exception of woman aged 15 to 19 years which is a reference group in the analysis consistent with the current practice in the Czech Republic. As the crucial date determining the age is set the 30^{th} of June 2010 for the analysis of the year 2011 and the 30^{th} of June 2011 for analysis of the year 2012. If the age is minus one, it is increased to zero.

5.1.3 Region

Regions are risk factors commonly used in both capitation models (e.g. in Switzerland or the Netherlands) and empirical works (Lamers and van Vliet, 2004; Huber et al., 2013). In our analysis, we will use the region of residence for each insuree in the detail of NUTS2[6]. The Czech Republic is divided into eight NUTS2 regions which can be seen in the Figure 5.3. The number of regions is comparable to the number of regions used in the risk model used in other countries e.g. in the Netherlands. Observations where the regional code was missing (n=2108, 0.2% of the sample) were dropped from the regression.

Figure 5.3: Czech Republic regions

Source: Regional council, edited by the author

The Table 5.2 shows the differences in mean health care costs among the regions in 2011. These differences again suggest that regions might be valid adjusters for health care costs. The highest average costs incurred the insurees living in Prague (22 479 CZK), whereas the individuals living in Southwest region had the lowest average cost (19 026 CZK).

Table 5.2: Average health care costs in the regions

Code	NUTS 2	Average costs	Population
CZ01	Prague	22 479	122 444
CZ02	Central Bohemia	20 302	127 974
CZ03	Southwest	19 026	121 144
CZ04	Northwest	19 031	113 592
CZ05	Northeast	19 593	152 430
CZ06	Southeast	19 129	169 283
CZ07	Central Moravia	19 129	123 587
CZ08	Moravia-Silesia	20 022	125 634

[6] The original data set included region for each insuree in the detail of districts (NUTS4). Including these 76 regional dummies might cause the model to be over-fitted, thus we decided to use NUTS2 regions.

5.1.4 Insurance companies

Table 5.3 displays the shares of population in 2012 registered with each insurance companies. The proportion which belonged to the companies is comparable to the values in our sample. Table 5.3 also summarizes the percentage of chronically ill people (classification to the PCGs based on 121DDD) and mean age, which are the leading factors of health care costs in our analysis. The insurance companies with a higher number of chronically ill people had higher expenditures on average. This fact confirms our hypothesis that a chronic condition based on past consumption of drugs might be a good risk adjuster. Nevertheless, this inference is not such straightforward. Each chronic condition represents different diseases, thus incurs a different health care cost.

Table 5.3: Population characteristics by insurance company

ID	**% in CR***	**% in our sample**	**% of chronic conditions**	**Mean age**	**Mean cost**
111	59.39%	60.09%	33.63%	43	21 507
201	6.01%	5.74%	28.07%	39	19 293
205	7.08%	6.98%	26.58%	37	17 934
207	6.71%	6.69%	26.32%	37	18 829
209	1.32%	1.29%	30.79%	39	19 286
211	11.37%	11.09%	27.93%	39	19 070
213	3.99%	4.09%	23.67%	36	16 927
217	4.13%	4.04%	25.17%	35	17 033

* http://www.mzcr.cz

Although the information about the insurance company which an individual belongs to is not necessary for calculating the risk adjustment model, it is important for evaluating the impact of the new model in practice. As mentioned earlier the improper risk adjustment model might cause financial problems to the insurance companies due to unfair allocation of money from the public insurance.

5.1.5 PCG

The consumption of specific drugs is an indicator for the presence of a severe disease. Only drugs prescribed specifically against certain chronic conditions can be used to assign persons to a PCG which are subsequently used as risk

adjusters. The codes from the prescriptions were first mapped to the WHO ATC codes. ATC drug classes labeled with the 3^{rd} to the 7^{th} level of ATC code are assigned to corresponding disease class (e.g. L02 Endocrine therapy is assigned to the group Hormonal oncological treatment). Each drug or drug class should only correspond to 1 disease category i.e. mutually exclusive.

To avoid the assignment of incidental users of drugs to chronic conditions, persons were assigned to a PCG based on certain threshold. The most commonly used threshold is the prescribed daily doses per year, but the exact value of this threshold varies. In our analysis, we will compare two models based on 121 DDD and 181 DDD which represent the drug consumption for three and six months. Each prescription available in the data set contains an ATC code which allows us to classify it to a PCG.

Our analysis is based on the PCG classification currently used in Slovakia (Ministry of Health, 2012) which we compared and further adjusted to the Dutch classification (Lamers and van Vliet, 2004).

The Slovakian classification is also based on the PCG classification system used in the study of Lamers and van Vliet (2004) (Table 2.1). Slovakian Ministry of Health excluded from this classification some ATC codes for drugs which are not available in Slovakia. On the other hand, in Slovakia some groups were added which were found to correspond better to the health status of the Slovakian population.

In our analysis, we kept updates made in PCG classification of Lamers and van Vliet (2004) in case we did not find enough observations of individuals classified in these chronic conditions i.e. Gout, Tuberculosis, and Peptic acid disease was excluded. The group of Rheumatologic conditions was divided into two subgroups which differs in the content of TNF inhibitors which treats mainly rheumatoid arthritis. Respiratory illness, asthma is divided into two groups of Severe asthma, Chronic Obstructive Pulmonary Disease (COPD) and Asthma, groups of Psychotic illness, Spinal cord and brain disease, Treatment with growth hormone, Hormonal oncological treatment and Neuropathic pain are set. We decided to keep Hypertension low, Hypertension high and Thyroid disorders as in Lamers and van Vliet (2004). The updated PCG classification which we use for our analysis is presented in the Table 5.4.

Table 5.4: PCG classification

PCG	Chronic condition	ATC codes	Exclusion	Prevalence 121DDD	Prevalence 181DDD
PCG1	Glaucoma	S01E		1.593%	1.293%
PCG2	Thyroid disorders	H03A,H03B		3.888%	2.625%
PCG3	Psychotic illness	N05A excluding (N05AL03, N05AN01), N06DA, N06DX01, N07BB, N07BC51		0.707%	0.538%
PCG4	Depression	N06A excluding (N06AA09, N06AX21)	if in PCG3	3.516%	2.892%
PCG5	Hyperlipidemia	C10 (excluding C10AC01, C10BX03)	if in PCG6, PCG9 or PCG16	7.919%	6.951%
PCG6	Diabetes with hypertension	A10&C02 excluding (C02KX, C02CA04) C03 excluding (C03CA01), C07, C08 excluding (C08CA06), C09		4.996%	4.961%
PCG7	Severe asthma, COPD	R03AC18, R03AK03, R03BB		0.483%	0.346%
PCG8	Asthma	R03 excluding (R03AC18, R03AK03, R03BB, R03CA02, R03BC01, R03CC02, R03CC13)	if in PCG7	2.638%	2.028%
PCG9	Diabetes type II	A10	if in PCG5, PCG6 or PCG16	0.438%	0.361%
PCG10	Epilepsy	N03 excluding (N03AX12, N03AX16, N03AE01)		0.656%	0.479%
PCG11	Crohn's and ulcerative colitis	A07EA06, A07EC02		0.223%	0.173%
PCG12	Cardiac disease	C01A, C01B, C01D, C01EB15, C01EB17, C03CA01		3.813%	3.253%
PCG13	Rheumatologic conditions	L04AA11, L04AA24, L04AB, L04AC		0.058%	0.053%
PCG14	Rheumatologic conditions II	A07EC01, L01BA01, L04AA13, L04AX03, M01CB01, M01CC01, P01BA02	if in PCG13	0.307%	0.183%
PCG15	Parkinson's disease	N04B		0.258%	0.210%
PCG16	Diabetes type I	A10A	if in PCG6	0.234%	0.225%
PCG17	Transplantations	L04AA06, L04AA10, L04AA18, L04AC02, L04AD01, L04AD02, L04AX01		0.149%	0.110%
PCG18	Cystic fibrosis	J01GB01, J01XB01, R05CB13		0.004%	0.003%
PCG19	Spinal cord and brain disease	L03AB07, L03AB08, L03AX13, L04AA23, M03BX01, M03BX02		0.101%	0.081%
PCG20	Malignancies	L01 excluding (L01BA01), L03AA, L03AC01, L04AX04		0.250%	0.184%
PCG21	HIV/AIDS	J05AE, J05AF excluding (J05AF08, J05AF10, J05AF11) J05AG, J05AR, J05AX excluding (J05AX05)		0.017%	0.015%
PCG22	Renal disease	B03X, V03AE		0.066%	0.054%
PCG23	Treatment with growth hormone	H01AC01, H01AC03		0.016%	0.015%

(Continued on next page)

Table 5.4: PCG classification (continued)

PCG	Chronic condition	ATC codes	Exclusion	Prevalence 121DDD	Prevalence 181DDD
PCG24	Hormonal oncolog. treatment	L02		0.371%	0.333%
PCG25	Neuropathic pain	N01BX04, N03AX12, N03AX16,		0.204%	0.112%
PCG26	Hypertension - low	C03A, C03EA01 C07	if in PCG6	5.422%	4.884%
PCG27	Hypertension - high	C02, C08, C09A, C09B	if in PCG6	8.930%	8.163%
Sum				**47.26%**	**40.53%**

Only the drugs mentioned in Table 5.4 are therefore indicative of chronic conditions which are subsequently used to predicts future spending of people classified in a category. The last two column of Table 5.4 present the prevalence of the conditions in our sample for the two threshold 121DDD and 181DDD. The most frequent disease in the population was Hypertension-high (more than 8%), followed by Hyperlipidemia and Hypertension - low. On the other hand, Cystic fibrosis, HIV/AIDS and Treatment with growth hormone are among the scarcest diseases.

Table 5.5 summarizes the number of PCGs per insuree.

Table 5.5: Number of PCGs per insuree

Number of PCGs	121DDD	181DDD
0	69.669%	72.553%
1	18.146%	17.519%
2	8.491%	7.372%
3	2.876%	2.099%
4	0.692%	0.399%
5	0.112%	0.054%
6	0.012%	0.004%
7	0.002%	0.000%
8	0.000%	-

An individual can be assign to one or more chronic conditions based on his drug consumptions i.e. if he suffers from Diabetes and Hypertension - high, he/she is classified in both PCGs, similar to the Netherlands. When the 121DDD threshold is set, almost 70% of people is not assigned to any PCG. For the 181DDD threshold, it rises to more than 72%. Those not assigned to any PCGs are categorized in a group "no PCG" and this group is used as the reference category in the model.

Chapter 6

Methodology

This chapter introduces the methodology used for the estimation and prediction of health care expenditures in the demographic and PCG models.

6.1 Process of estimation

To assess the accuracy of the models for predicting future costs, the second year of the sample is not included in the estimation of 6.1, 6.3 and 6.5 and is used for comparison of the predictive power of estimated models. The process follows in three steps:

1. models are fit to the estimation data set (year t)
2. estimated coefficients are, then, used to calculate predicted cost in year t+1
3. the predicted costs are compared with the actual costs

We will compare the financial impact on individual insurance companies resulting from the demographic and PCG models in 2012. Since we do not have data on money allocated to each insurance company, the financial impact will be approximated based on the share of insurees in 2012. We will use the real health care costs incurred for our sample in 2012 as the amount for redistribution among the insurance companies.

6.2 Determination of the models

In all models included in our analysis, the dependent variable is the individual health care expenditure, and the independent variables are dummy variables defining different risk classes in the population. As mentioned in the Chapter 3.2 the insurance companies receive state payment for an insured each month. Nevertheless, the cost indices are calculated once a year (in year t for year t+1). These indices are then used to calculate monthly capitation payment allocated to insurance companies. In our study we will keep the same method of estimation, thus the models will be calculated on annual basis.

We compare three basic models.

6.2.1 Demographic model

The first type of model to be estimated is the model currently used in the Czech Republic. It predicts health care expenditures based on the categorization of each individual to age/sex groups. The demographic model is defined as:

$$cost_i = \alpha + \beta_{sa} sex\_age_i, \tag{6.1}$$

where $cost_i$ are health care costs for the i^{th} person and sex_age_i is the sa^{th} age/sex group for the i^{th} person. When the coefficients of the model are estimated, expenditures for the next year will be predicted and compared to real historical data.

Predicted cost for the demographic model (equation 6.1) for the year t+1 are given by:

$$\hat{cost}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_{sa,t} sex\_age_{i,t+1}, \tag{6.2}$$

where the coefficients $\hat{\alpha}_t$ and $\hat{\beta}_{sa,t}$ are estimated by equation 6.1. The demographic variables for the prediction of cost for year t+1 is also based on the individual age in t+1 rather than t. As the insurance companies submit the demographic data monthly to the Ministry of Health, it is possible to use the current age structure to take advantage of the most up-to-date data in the model e.g. a male 49 years old in year t was classified to the group of males aged 45-49 years for the analysis of year t, for the prediction of health care costs

in year t+1 he is classified to the group of 50-54 as he is 50 years old in year t+1.

6.2.2 PCG model

The second model includes the same risk adjusters as equation 6.1 and adds dummy variables indicating the presence of a chronic condition (PCG) based on drug utilization in year t. It is defined as:

$$cost_t = \alpha + \beta_{sa} sex\_age_i + \gamma_x pcg\_x_i, \tag{6.3}$$

where $cost_i$ are health care costs for the i^{th} person, sex_age_i is the sa^{th} age/sex group for the i^{th} person and pcg_x is the x^{th} PCG group for the i^{th} person.

Predicted expenditures for the PCG model are determined as:

$$\hat{cost}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_{sa,t} sex\_age_{i,t+1} + \hat{\gamma}_{x,t} pcg\_x_{i,t}, \tag{6.4}$$

where the coefficients $\hat{\alpha}_t$, $\hat{\beta}_{sa,t}$ and $\hat{\gamma}_{x,t}$ are estimated by Equation 6.3. We use again the individual age for the year t+1, contrary to 6.3 the PCG classification is based on the consumption from the year t e.g. an individual who was classified to a PCG in year t is also classified to the same PCG for the prediction of health care cost in t+1. The demographic variables for the prediction of cost for year t+1 are also based on the individual age in t+1, similar to Equation 6.2.

6.2.3 PCG and region model

The third model comprises the variables from equation 6.3 and adds regional dummies. It is defined as:

$$cost_t = \alpha + \beta_{sa} sex\_age_i + \gamma_x pcg\_x_i + \delta_r region_i, \tag{6.5}$$

where $cost_i$ are health care costs for the i^{th} person, sex_age_i is the sa^{th} age/sex group for the i^{th} person, $region_i$ is the r^{th} region of residence for the i^{th} person and pcg_x is the x^{th} PCG group for the i^{th} person.

Predicted expenditures for the PCG and regional model (6.5) are determined as:

$$\hat{cost}_{i,t+1} = \hat{\alpha}_t + \hat{\beta}_{sa,t} sex\_age_{i,t+1} + \hat{\gamma}_{r,t} region_{i,t+1} + \hat{\delta}_{x,t} pcg\_x_{i,t}, \tag{6.6}$$

where the coefficients $\hat{\alpha}_t$, $\hat{\beta}_{sa,t}$ and $\hat{\gamma}_{r,t}$ are estimated by Equation 6.5. We use again the individual age and also the region of residence for the year t+1 as opposed to 6.5.

All models will be estimated using the OLS estimation approach on untransformed data. All models are therefore assumed to be linear in the coefficients and include an intercept. They are estimated by means of WLS where the regression weights are a function of individual eligibility in the given period. The prediction which best reflects the reality will be determined as the most efficient model from equations 6.1, 6.3 and 6.5. As one of the objectives of our analysis is to develop a method of risk adjustment that might be put into practice by the regulators and policymakers, the advantage of the WLS models is that they are easier to use and to interpret than other models.

6.3 Model evaluation

Efficiency of a particular model is based on its predictive performance at the population as a whole evaluated by the three measures described in sections 6.3.1-6.3.4. Additionally, we examine how well the model predicts health care costs for different cost sub-groups in the population.

6.3.1 R^2

The parameter R^2 measures the percent of individual variance explained by the model and indicates how much of the variance in the dependent variable is explained by the model. The R^2 of the OLS regression is defined as:

$$R^2 = \frac{SSE}{SST}, \tag{6.7}$$

where SSE is the explained variation and SST is the total variation. The higher

the R^2 value, the higher the predictive performance of the model. The parameter R^2 will evaluated the models at the population level.

6.3.2 MAPE

Mean Absolute Prediction Error (MAPE) is calculated as the average of the absolute differences between predicted and observed health care costs (Van de Ven et al., 2014):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}|C_i - \hat{C}_i| \tag{6.8}$$

where C_i is the actual costs, $\hat{C}_i$ the costs predicted by the model and n is the number of individuals in the sample. Using the absolute value means that predictions that are greater or less than actual costs cannot cancel each other out. Lower MAPE-values indicate a higher predictive performance of the model.

6.3.3 MARE

Mean Absolute Relative Error (MARE) captures the relative accuracy of a prediction and is defined as the average of the absolute differences between predicted and observed health care costs divided by the actual cost:

$$MARE = \frac{1}{n}\sum_{i=1}^{n}\frac{|C_i - \hat{C}_i|}{C_i} \tag{6.9}$$

Because some subjects have no costs in the forecast year, we add a value of 1 to actual costs for all subjects to avoid undefined ratios. The MARE is a relative measure that expresses errors as a percentage of the actual data. This is its greatest advantage as it provides an easy and intuitive way of judging the extent or importance of errors. As it is a relative number without dimensions, the comparison of the model is straightforward. A risk instrument's forecast power is evaluated by the degree its MARE closer to 0. The lower MARE the model has, the more accurate the model is.

6.3.4 Cost quintile analysis

Risk instruments may generate more accurate predictions for different ranges of the cost distribution. We evaluate the relative strength of the PCG model to predict costs for relatively high, medium, and low-cost subjects by examining the predictive performance of the model by cost quintile. Subjects are grouped into five equally populated segments based on their actual costs, and the values of MAPE and MARE are compared. The MAPE and MARE measures survey subgroups and can provide a good indication of the extent to which models compensate insurers for differences in expenses between cost subgroups. Certain models may overpredict low-cost groups or underpredict high-cost groups more significantly. This method is also applied in studies Fishman and Shay (1999); Fishman et al. (2003); Van de Ven et al. (2014).

Chapter 7

Results

This chapter discusses the results of our analysis. Firstly, it comments and compares the results obtained for each risk adjustment model. Secondly, the impact of selected model on the insurance companies is calculated.

7.1 Model estimation

In this section, we present the results of models 6.1-6.6 that we analyzed in detail. Table 7.1 summarizes the variables included in each model and the DDD threshold for classifying a patient to a PCG group. We assigned a number 1-6 to each model to better distinguish between them when displaying their results.

Table 7.1: Description of the estimated models

Model number	**Equation number**	**Name**	**Dummy variables**	**DDD threshold**
Model 1	6.1 & 6.2	Demographic model	36 age/sex groups	-
Model 2	6.1 & 6.2	Demographic model	38 age/sex groups (age 0 added)	-
Model 3	6.3 & 6.4	PCG	38 age/sex, 27 PCG groups	181DDD
Model 4	6.5 & 6.6	PCG+regions	38 age/sex, 27 PCG, 8 region groups	181DDD
Model 5	6.3 & 6.4	PCG	38 age/sex, 27 PCG groups	121DDD
Model 6	6.5 & 6.6	PCG+regions	38 age/sex, 27 PCG, 8 region groups	121DDD

The number of variables included in the model for age/sex and region groups is always one less than the number of defined risk classes because of perfect multicollinearity if otherwise. All 27 PCG groups are included in the regressions as their reference category is no PCG, the case when an individual was not assigned to any chronic condition based on one's past consumption.

The estimation results of the models for the year t (2011) are displayed in Table 7.2.

Table 7.2: Estimation results - validation period 2011

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
M 0y		32 746 **	33 712 **	33 581 **	33 991 **	33 859 **
M 1-4	6 032 **	-98	510	430	706	624
M 5-9	-796 **	-796	-712	-757	-585	-632
M 10-14	-672	-672	-847	-851	-826	-831
M 15-19	-2 220	-2 220 **	-2 233 **	-2 240 **	-2 152 **	-2 159 **
M 20-24	-3 279 **	-3 279 **	-3 631 **	-3 673 **	-3 507 **	-3 551 **
M 25-29	-2 393 **	-2 393 **	-3 203 **	-3 294 **	-3 105 **	-3 198 **
M 30-34	-1 371 **	-1 371 **	-2 601 **	-2 718 **	-2 525 **	-2 643 **
M 35-39	-371 **	-371	-2 162 **	-2 251 **	-2 252 **	-2 342 **
M 40-44	1 614	1 614 **	-872	-956	-937	-1 023
M 45-49	4 279 **	4 279 **	289	219	191	119
M 50-54	9 136 **	9 136 **	3 021 **	2 974 **	2 575 **	2 524 **
M 55-59	17 066 **	17 066 **	7 254 **	7 184 **	6 542 **	6 469 **
M 60-64	24 681 **	24 681 **	11 136 **	11 050 **	10 057 **	9 966 **
M 65-69	33 013 **	33 013 **	15 637 **	15 516 **	14 136 **	14 010 **
M 70-74	39 537 **	39 537 **	18 825 **	18 724 **	16 435 **	16 330 **
M 75-79	44 849 **	44 849 **	21 515 **	21 407 **	18 727 **	18 613 **
M 80-85	47 738 **	47 738 **	23 385 **	23 246 **	20 397 **	20 251 **
M 85up	48 191 **	48 191 **	24 894 **	24 670 **	22 292 **	22 059 **
F 0y		30 407 **	31 447 **	31 315 **	31 724 **	31 590 **
F 1-4	5 035 **	-951	-184	-260	29	-48
F 5-9	-2 534 **	-2 534 **	-2 281 **	-2 332 **	-2 074 **	-2 126 **
F 10-14	-1 131 **	-1 131	-1 210	-1 223	-1 156	-1 170
F 20-24	-457	-457	-773	-802	-748	-777
F 25-29	2 612	2 612 **	1 660 **	1 560 **	1 633 **	1 532 **
F 30-34	4 020 **	4 020 **	2 500 **	2 377 **	2 408 **	2 284 **
F 35-39	3 637 **	3 637 **	1 281 **	1 194 **	1 158 **	1 070
F 40-44	4 100 **	4 100 **	644	576	400	332
F 45-49	6 919 **	6 919 **	1 799 **	1 723 **	1 370 **	1 293 **
F 50-54	10 360 **	10 360 **	3 088 **	3 043 **	2 386 **	2 340 **
F 55-59	13 956 **	13 956 **	3 829 **	3 757 **	2 817 **	2 743 **
F 60-64	18 458 **	18 458 **	5 517 **	5 422 **	4 159 **	4 063 **
F 65-69	24 575 **	24 575 **	7 584 **	7 460 **	5 997 **	5 871 **
F 70-74	30 651 **	30 651 **	10 756 **	10 687 **	8 685 **	8 615 **
F 75-79	36 735 **	36 735 **	13 883 **	13 796 **	11 292 **	11 201 **
F 80-85	39 512 **	39 512 **	15 545 **	15 404 **	12 610 **	12 464 **
F 85up	41 745 **	41 745 **	18 917 **	18 749 **	15 839 **	15 664 **
Intercept	10 330 **	10 330 **	9 269 **	8 683 **	8 967 **	8 453 **
PCG 1			4 896 **	4 780 **	5 264 **	5 162 **

(Continued on next page)

Table 7.2: Estimation results (Continued)

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
PCG 2			4 415 **	4 340 **	3 853 **	3 788 **
PCG 3			46 249 **	46 279 **	43 031 **	43 055 **
PCG 4			20 165 **	20 079 **	19 694 **	19 609 **
PCG 5			7 498 **	7 431 **	7 178 **	7 104 **
PCG 6			23 564 **	23 654 **	22 927 **	23 019 **
PCG 7			38 711 **	38 690 **	34 201 **	34 194 **
PCG 8			16 810 **	16 776 **	15 234 **	15 202 **
PCG 9			7 202 **	7 235 **	7 186 **	7 219 **
PCG 10			31 286 **	31 327 **	28 303 **	28 340 **
PCG 11			14 571 **	14 580 **	10 299 **	10 306 **
PCG 12			35 614 **	35 704 **	34 021 **	34 110 **
PCG 13			338 557 **	338 411 **	324 573 **	324 424 **
PCG 14			39 802 **	39 863 **	29 561 **	29 651 **
PCG 15			29 923 **	29 932 **	24 418 **	24 424 **
PCG 16			45 802 **	45 862 **	45 548 **	45 605 **
PCG 17			135 265 **	135 241 **	119 484 **	119 462 **
PCG 18			448 431 **	448 411 **	421 751 **	421 813 **
PCG 19			209 259 **	209 207 **	186 144 **	186 106 **
PCG 20			441 534 **	441 511 **	380 126 **	380 110 **
PCG 21			206 401 **	205 832 **	196 849 **	196 287 **
PCG 22			848 523 **	848 465 **	807 401 **	807 340 **
PCG 23			238 110 **	238 094 **	225 440 **	225 438 **
PCG 24			49 050 **	48 986 **	48 334 **	48 265 **
PCG 25			53 015 **	52 955 **	46 686 **	46 662 **
PCG 26			4 880 **	4 939 **	4 513 **	4 579 **
PCG 27			3 712 **	3 786 **	3 641 **	3 724 **
CZ01				3 170 **		3 099 **
CZ02				1 161 **		1 097 **
CZ04				175		83
CZ05				360		283
CZ06				-306		-398
CZ07				-88		-147
CZ08				1 208 **		1 104 **
R^2	0.0207	0.0283	0.1851	0.1853	0.1947	0.1948

Note: Significance codes * p-value< 0.1 ; ** p-value < 0.05

The R^2 of the models predicting the total cost for year t ranged from 0.0207 to 0.1948. In the models which control for demographic measures, the R^2 was the lowest: 0.0207 for the model used currently and 0.0283 for the model when the groups of newborn (up to one year) male/female were added. The PCG models reported higher R^2. The model based on 181DDD reached the R^2 of 0.1851. When the region dummies added, the R^2 rose only by 0.02 to 0.1853. A similar result was found for the PCG model based on 121DDD threshold which reported the R^2 of 0.1947 and 0.1948 for the PCG region model.

Not all the age groups turned out to be significant in the demographic, nor the

PCG models (for Model 3 to 6 the insignificant groups were males aged 1-4, 10-14, 40-44 and 45-49 years and females aged 1-4, 10-14, 20-24, 35-39 and 40-44 years). Therefore, we tested models with age groups defined differently. We clustered the insignificant groups into 10-20 years instead of 5 years, but all the models estimated costs in year t as well as the predictive models for expenditures in year t+1 performed worse. Thus, we kept the logic of the demographic model currently used and age variables in all models are grouped by five years except the significant group of newborns. All of the PCG regression coefficients are significantly different from zero (p-value=0.000). Only three regions' coefficients turned out to be significant with p-value<0.05. Moreover, the contribution of adding the region variables to the PCG model is low. The R^2 rose only by 0.01 and 0.02 for PCG model based on 181DDD and 121DDD, respectively. It suggests the region variables are not valid risk adjusters of health care costs in the Czech Republic.

The intercept of a model represents the health care cost for the reference groups of the model. i.e in PCG model a women aged 15-19 years and classified in no PCG. The highest intercept is reached for the demographic models (10 330 CZK), the lowest values are reported by the region PCG models based on 121DDD (8 452 CZK). As all variables are dummy variables, their coefficients represent the cost of a patient who is classified (dummy=1) to demographic, region or PCG group. When the coefficients for risk groups are reported, we need to consider the value of intercept as well e.g. for a male aged 20-24 years the average cost is 10 330 - 3 279 = 7 051 CZK. The most expensive individuals with regard to age are males 85 years and older. Their health care cost are estimated at more than 48 000 CZK for demographic models, at 24 000 CZK for PCG model based on 181DDD and more than 22 000 CZK for the PCG model based on 121DDD. The least expensive group are the males 20-24 years old which cost less than -3 000 CZK.

Among all chronic conditions, Renal disease (PCG 22) is associated with the highest health care costs (more than 807 000 CZK for all models), following by Cystic fibrosis (PCG 18) with more than 420 000 CZK and Malignancies (PCG 20) incurring health care cost of more than 440 000CZK for Model 3 and 4 and 380 000 CZK for Model 5 and 6. Among the chronic groups that incur up to 5 000 CZK are Hypertension - high (PCG 27), Thyroid disorders (PCG 2) and Glaucoma (PCG 1). The coefficients of PCG are generally higher in the model based on 181DDD as only patients with more severe condition (higher number

of DDD) are classified to a chronic group.

In our analysis, the main criterion for evaluation of the models is how well it predicts the future health care cost, thus we focus on the predictive power of the models for the year t+1 (2012). The predictive performance was assessed and compared at both the population and the cost quintile level. By doing so, it is possible to examine how well the models predict expenses for the total sample and different cost sub-populations.

Table 7.3 shows the R^2, MAE and MARE for the prediction models of health care expenditures in year t+1.

Table 7.3: Evaluation of models for year t+1

	R^2	**MAPE**	**(SD)**	**MARE**	**(SD)**
Model 1	0.0203	22 709	(118 704)	178.88	(1 929)
Model 2	0.0206	23 227	(118 933)	176.65	(2 004)
Model 3	0.1295	19 856	(115 934)	117.48	(1 295)
Model 4	0.1297	19 853	(115 929)	122.78	(1 347)
Model 5	0.1387	19 775	(115 685)	111.61	(1 229)
Model 6	0.1389	19 770	(115 680)	116.94	(1 280)

The models for the year t+1 had lower predictive R^2 than the prediction models for year t as supposed. No more than 14% of the variations in the total cost for t+1 was explained. Although these R^2 values might seem low, they are consistent with empirics. Van Vliet (1992) concluded that no more than 20% of the variations are predictable on an individual basis. Fishman et al. (2003) argues that the maximum explainable R^2 for risk assessment models may be approximately 30%. Risk-based forecast produced relatively low R^2 because a large portion of health care costs occur randomly and cannot be predicted with great confidence. There are outliers with extremely high costs in our sample which are not predicted adequately by the risk adjusters and therefore lead to lower R^2. The high-cost patients are controlled for only when they classify to a PCG. Around 70% of individuals do not classify in any PCG (see Table 5.5) and for those individuals we do not have any other health status proxy which would capture their health conditions. We suppose them to be in "good health" and they are risk adjusted only by demographic variables.

Demographic risk instrument explains relatively little of the variance in cost in year t+1 with an R^2 of 0.0206. Poor predictive power of these models is also confirmed by large values of MAPE and MARE. R^2 increases to almost 0.13

and the MAPE decreases when the PCG groups are added with the threshold of 181DDD. When the threshold value for the PCG model is set to 121DDD, a remarkable improvement in the predictive power is reached: R^2 of 0.1387, MAPE 19 775 CZK and the MARE of 111.61. The results also confirmed that adding the region risk adjusters does not improve considerably the predictive power of the PCG models. The R^2 value rose only by 0.02 when the region variables added to the PCG models.

7.2 Cost quintile analysis

Table 7.4 shows model performance for five quintiles of health care cost in 2012. The first row for each quintile represents the mean actual value of health care costs (in CZK) for the year 2012 and its standard deviation for this quintile. Then the predicted mean health care cost per quintile is calculated for each model as well as the values for MAPE and MARE. We consider a prediction model more efficient when the mean per quintile is closer to the actual mean value and again the values of MAPE and MARE are low (the lower, the better).

The demographic models tend to over-predict low cost (quintiles 1-4) and under-predict high cost (quintile 5) enrollees significantly. The reason is demographic models have larger intercepts and higher coefficients for the demographic groups than the PCG models (see Table 7.2) resulting in over-prediction of low cost and under-prediction of high-cost enrollees. The risk adjustment for demographic models is established only through means of age and sex and does not take into account other factors which determine the health care cost. Chronic conditions are therefore more accurate risk instrument when we evaluate predicted mean for the quintiles.

When comparing the PCG models, PCG model based on 121DDD predicts mean health care expenditures better than PCG models based on 181DDD with few exceptions e.g. for the prediction of mean for the fourth quintile, but the difference is small (17 CZK difference in means between the models 3 and 5). The greatest difference among PCG models occurs in the most expensive cost quintile where the mean prediction by the PCG model based on 181DDD is 41 789CZK whereas 43 011 for the PCG model based on 121DDD. On the other hand, when models are compared relatively in the 5th quintile PCG model (181DDD) turns out be by 0.02 more accurate.

Table 7.4: Model performance: Cost quintile analysis

Quintile	Model	Mean	(SD)	MAPE	(SD)	MARE	(SD)
1st	Actual value	1 223	(602)				
	Model 1	13 918	(8 712)	12 695	(8 710)	887.53	(4 240)
	Model 2	13 936	(9 424)	12 713	(9 420)	876.42	(4 412)
	Model 3	9 755	(5 064)	8 529	(5 060)	583.43	(2 853)
	Model 4	9 766	(5 180)	8 539	(5 187)	609.99	(2 966)
	Model 5	9 349	(5 008)	8 122	(5 003)	554.12	(2 706)
	Model 6	9 359	(5 118)	8 133	(5 123)	580.84	(2 818)
2nd	Actual value	3 322	(666)				
	Model 1	15 261	(9 214)	11 939	(9 181)	3.74	(2.97)
	Model 2	14 829	(10 442)	11 507	(10 421)	3.61	(3.33)
	Model 3	10 983	(6 184)	7 661	(6 158)	2.41	(1.98)
	Model 4	10 934	(6 259)	7 612	(6 233)	2.39	(2.01)
	Model 5	10 653	(6 387)	7 331	(6 358)	2.30	(2.02)
	Model 6	10 604	(6 454)	7 282	(6 426)	2.29	(2.04)
3rd	Actual value	6 485	(1 255)				
	Model 1	18 744	(11 603)	12 279	(11 517)	1.97	(1.87)
	Model 2	18 495	(13 908)	12 030	(13 815)	1.92	(2.2)
	Model 3	14 467	(11 016)	8 111	(10 818)	1.29	(1.67)
	Model 4	14 447	(11 037)	8 135	(10 806)	1.29	(1.67)
	Model 5	14 220	(11 445)	7 894	(11 232)	1.25	(1.74)
	Model 6	14 200	(11 458)	7 923	(11 210)	1.25	(1.74)
4th	Actual value	13 548	(3 248)				
	Model 1	23 818	(14 132)	12 566	(12 177)	0.98	(1.03)
	Model 2	24 448	(18 201)	13 574	(16 305)	1.05	(1.33)
	Model 3	21 910	(19 731)	12 133	(17 499)	0.92	(1.35)
	Model 4	21 921	(19 731)	12 163	(17 482)	0.92	(1.35)
	Model 5	21 927	(20 648)	12 365	(18 351)	0.93	(1.4)
	Model 6	21 936	(20 637)	12 390	(18 324)	0.93	(1.4)
5th	Actual value	87 152	(262 670)				
	Model 1	30 238	(15 718)	64 067	(260 521)	0.54	(0.3)
	Model 2	30 998	(22 523)	66 308	(260 282)	0.61	(0.44)
	Model 3	41 789	(65 966)	62 824	(253 696)	0.64	(0.67)
	Model 4	41 837	(65 962)	62 795	(253 689)	0.64	(0.67)
	Model 5	43 011	(67 685)	63 138	(252 955)	0.66	(0.71)
	Model 6	43 056	(67 674)	63 100	(252 949)	0.66	(0.71)

Thus, generally the differences in the prediction accuracy for the cost quintiles between 121DDD and 181DDD PCG models are small, but the PCG model of 121DDD slightly outperform the other models. We can also conclude that adding the region adjusters does not bring any significant improvement in the prediction accuracy, neither in the overall nor quintile analysis.

7.3 Impact on the insurance funds

Table 7.5 shows the financial impact of PCG model (121DDD) implementation on the insurance companies when compared to the allocation of the amount by the current demographic model.

Table 7.5: Financial impact of PCG model implementation

ID	Revenues difference (CZK)	Relative revenues
111	69 413 395	100.46%
201	-3 012 622	99.76%
205	-24 153 087	98.35%
207	-18 783 776	98.64%
209	7 654 229	102.68%
211	-1 241 633	99.95%
213	-23 155 784	97.13%
217	-6 720 720	99.15%

The greatest impact in terms of CZK would be on Všeobecná zdravotní pojišťovma (VZP - 111) which would receive more than 69 million (0.46%). Zaměstnanecká pojištovna Škoda (209) would also improve its financial position by more than 7 million. The rest of the insurance companies would earn less, the most substantial loss of 23 million (-2.87%) would suffer Revírská bratrská pokladna (213).

In recent years, VZP suffered from financial problems. It is due to its distribution of insurees who are older and more ill compared to other insurance companies (see Table 5.3). In 2013, the Ministry of Health granted an interest-free loan to VZP amounting 1.7 billion CZK. The first part (0.7 bil.) was repaid last year; the rest should be pay off in years 2015 and 2016. Therefore increasing its revenues systematically should be the intention of our political representation. The establishment of the PCG model as the risk adjustment model is one of the possibilities how to allocate the financial resources more equitably among the insurance companies and therefore prevent their financial problems.

Chapter 8

Discussion

Our main contribution is that we confirmed that PCG model is a valid risk adjuster for individual health status in the Czech Republic. An increase of R^2 from 0.0203 for the demographic model currently used to 0.1387 for the PCG model based on 121DDD is significant improvement in model performance. We analyzed the performance of models on a large sample; therefore inferences from our analysis can be assumed to be valid for the whole population as well. We suggested a model that is easy to implement in reality and could be use by the Ministry of Health for risk adjustment in the Czech Republic.

The empirical analysis and the data used illustrated the potential for improving the predictive performance of the risk adjustment model in health care. The results of our analysis confirm that the PCGs contribute significantly to the prediction of individual expenditures. The main reason PCG model outperforms the demographic model is that PCG model uses the information on individual health condition to create risk categories. The demographic model assumes that everybody of the same age and sex incurs similar health care expenditures, which does not hold in reality.

Adding PCGs to the demographic model brings additional details about individual health status. When they are not assigned to any PCG group, they are supposed to be in good health and vice versa. We report in Table 5.5 that about 30% of enrollees in our sample (depending on the DDD threshold) classify into a PCG category. Any such information about health condition except sex and age is missing in the demographic model.

PCG model with regional variables does not increase the predictive power of the PCG model. As mentioned in the Chapter 4.1.3, the region information was not available for all cases. Statistical significance was not convincing neither; thus we do not believe the region risk classes should be used in the model for risk adjustment in the Czech Republic. One of the reasons that the region adjusters do not play a role in the models might be that people have official permanent address to the actual place of residence. The regional health differences are probably captured by demographic and morbidity variables as well. Therefore, the regional variables turn out to be insignificant.

For the case of PCG groups, a statistical significance should not be the exclusive condition about a model evaluation. When risk adjustment is based on previous drug consumption, there is always the danger of adverse selection and of rewarding inefficient providers. The reason is that the providers might be encouraged to more drug utilization than is strictly necessary (Lamers and Vliet, 2003). An individual who is assigned to one or more of the PCGs gives an advantage to the insurance fund to receive higher capitation payment, therefore in theory it would be financially advantageous for the fund to stimulate the prescription of drugs that fall into the PCG system. The Ministry of Health initiated projects in last years for the reduction of expenditures on drug prescription e.g. they supported the prescription of drugs with lower cost when generic substitutes (Ministerstvo zdravotnictvi CR, 2007). These initiatives to reduce costs of prescribed drugs while maintaining the quality of care should not be spoiled by contrary incentives of the risk adjusted capitation payments system.

To prevent gaming in prescription of drugs, we used the number of DDD instead of number of prescriptions which was also used in earlier studies (Lamers and Vliet, 2003; Prinsze and van Vliet, 2007). Using DDD eliminates the possible form of manipulation in this context e.g. a patient could receive four times a prescription for 1 week instead of a prescription for whole month. When the results of PCG models based on 121DDD and 181DDD in terms of R^2 differs only by 1%, we might also consider the threshold of 181DDD. A higher threshold is less prone to gaming as it is harder to reach it by stimulation of prescriptions. No risk assessment model can eliminate gaming, but every risk model should at least increase the cost of gaming. Gaming a PCG model by dispensing drugs that are not necessarily medically indicated involves costs to both the health insurance company and patient (if copayments apply) and also potential

health risks to the patient. As 121DDD represent a prescription of drugs for four months, we think it is sufficient level for the risk-adjustment model to prevent gaming.

When considering the application of the PCG model in reality, its important advantage is simple implementation. Cost and drug consumption information from previous years is, in most situations, already available in the administrative files of insurers and Ministry of Health. This means that it does not require a large additional administrative burden for collecting this information. Thus, regulators and policymakers could relatively easily improve the predictive performance of currently used models by including chronic condition information based on the drug consumption from prior years.

Chapter 9

Conclusion

Czech Ministry of Health is currently modeling the risk costs of insurance companies based on a demographic model. Such model takes into account demographic characteristics of the enrollees only, i.e. age and gender. Unfortunately, the demographic model does not yield much convincing results, achieving the explanatory power of 2%. This causes inefficiencies in government funds' allocation among insurance companies in the Czech Republic, since the model underestimates the costs of insurers with a large proportion of chronically ill enrollees. This leads to risk selection - insurers prefer to insure rather healthy patients and receive the same funds.

Recent research and experience from abroad (Netherlands, Slovakia) have shown that adding pharmacy-based cost groups adjusters into the model significantly improves its predictive power and would thus mitigate government funds allocation inefficiencies. Moreover, adding so-called regional risk adjusters could further upgrade the health care risk adjustment model in use. However, health care systems across countries tend to be very specific and the same frameworks might thus not be applicable globally. On contrary, each risk cost model's adjustment needs to be tested for carefully, taking into consideration local specifics.

The aim of this thesis was to test for PCG and regional risk adjusters in the context of the Czech health care system. Our analysis has shown that PCG adjusters notably upgrade the existing risk costs model, since they are able to account for the group of pharmaceuticals each individual is prescribed. This in turn makes the health care risk costs assessment more accurate and

helps improve funds' allocation to insurance companies. Therefore, we strongly recommend to include these PCG adjusters into the model of the Ministry of Health of the Czech Republic. Regional risk adjusters could further improve the model as they control for the region the enrollee comes from and thus account for region-specific effects on the health care costs, e.g. air pollution. Quite surprisingly, regional risk adjusters did not help improving the model's power and are thus seen as redundant. We suppose that the regional specifics are already captured in the individual health status defined by PCG.

The performance of the models was tested using a sample of 1 058 197 observations representing 10% of the Czech population in 2011 and 2012. The PCG classification included 27 chronic conditions. The models were estimated by WLS where the weight was a function of individual eligibility in the estimation period. Three types of model were assessed: demographic model, PCG model and PCG including region adjusters. Furthermore, two thresholds of 121 DDD and 181 DDD for classifying to the PCG were set and the performance of the PCG models was mutually compared. The model performance for predicting the health expenditures in the next year was evaluated by R^2 parameter. Its value increased from 2.03% for the demographic model currently used to 0.0206% when the age group of newborns was added to the model. The PCG models performed significantly better. The R^2 of PCG model based on 121DDD increased to 13.87%, and when the threshold of 181 DDD was set, the R^2 decreased little to 12.95%. The higher threshold for the PCG means that fewer people classify to a chronic condition. As the R^2 of the model is higher, the best model for predicting the health care expenditures is the PCG model with 121 DDD threshold.

These high R^2 values demonstrate that PCG model is able to predict more accurately individual health care costs than demographic model. We concluded that adding the PCG variables to the risk adjustment model increases its predictive performance which is in line with conclusions of other studies.Quite surprisingly, the regional risk adjusters did not bring any significant improvement of the model performance. The R^2 value for the PCG model including regional adjusters was 12.97% for the 181 DDD threshold and 13.89% for the 121 DDD threshold, almost equal values to PCG model. Thus, we do not recommend to include region adjusters into the risk adjustment model.

This thesis makes three important contributions to the literature dealing with

PCG models. First, we updated the PCG classification of Lamers and van Vliet (2004) by including new disease categories which are appropriate to capture morbidity conditions for the Czech population. Second, we added two new demographic groups, male and female newborns, as their actual mean health care costs are significantly different from the other children up to 5 years old. Third, we assessed the predictive performance of the PCG model based on 121 DDD and 181 DDD threshold and compared the results to the demographic model currently used in the Czech Republic.

There are several potential improvements to the risk adjustment model in the Czech Republic that are worth further research. It is expected that cost and diagnostic information from multiple prior years could further improve models' predictive performance. Therefore, the definition and classification of diagnostic groups could be investigated. Moreover, the number of years of lagged cost and diagnostic information could be explored to find what is the relevant period to be included in the risk adjustment model.

Bibliography

Ash, A. S., Porell, F. W., Gruenberg, L., Sawitz, E., and Beiser, A. (1990). Adjusting medicare capitation payments using prior hospitalization data.

Ash, H. B. A. S. (1997). *Diagnosis-based risk adjustment for Medicare capitation payments*. PhD thesis, University of Massachusetts Medical School.

Barros, P. P. (2003). Cream-skimming, incentives for efficiency and payment system. *Journal of Health Economics*, 22(3):419–443.

Buchner, F., Goepffarth, D., and Wasem, J. (2013). The new risk adjustment formula in germany: implementation and first experiences. *Health Policy*, 109(3):253–262.

Culyer, A. J. and Newhouse, J. P. (2000). *Handbook of health economics*. Elsevier.

Dunn, G., Mirandola, M., Amaddeo, F., and Tansella, M. (2003). Describing, explaining or predicting mental health care costs: a guide to regression models methodological review. *The British Journal of Psychiatry*, 183(5):398–404.

Eijkenaar, F. (2013). *Pay-for-performance for healthcare providers: Design, performance measurement, and (unintended) effects*. Instituut Beleid en Management Gezondheidszorg (iBMG).

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27.

Ellis, R. P. (2008). Risk adjustment in health care markets: concepts and applications. *International Health Care Financing*.

Ellis, R. P., Ash, A., et al. (1988). *Refining the Diagnostic Cost Group Model: A Proposed Modification to the AAPCC for HMO Reimbursement*. Brandeis Health Policy Research Consortium, Boston University.

Ellis, R. P., Fiebig, D. G., Johar, M., Jones, G., and Savage, E. (2013). Explaining health care expenditure variation: Large-sample evidence using linked survey and health administrative data. *Health economics*, 22(9):1093–1110.

EU (2012). Directive n. 442/2012 sb.

Fishman, P. A., Goodman, M. J., Hornbrook, M. C., Meenan, R. T., Bachman, D. J., and Rosetti, M. C. (2003). Risk adjustment using automated ambulatory pharmacy data: the rxrisk model. *Medical care*, 41(1):84–99.

Fishman, P. A. and Shay, D. K. (1999). Development and estimation of a pediatric chronic disease score using automated pharmacy data. *Medical care*, 37(9):874–883.

Greene, W. (1997). *Econometric Analysis*. Prentice-Hall international editions. Prentice Hall.

Huber, C. A., Schneeweiss, S., Signorell, A., and Reich, O. (2013). Improved prediction of medical expenditures and health care utilization using an updated chronic disease score and claims data. *Journal of clinical epidemiology*, 66(10):1118–1127.

Jones, A. M. (2010). Models for health care. *HEDG Working Paper 10/01*.

Keenan, P. S., Buntin, M. J. B., McGuire, T. G., and Newhouse, J. P. (2001). The prevalence of formal risk adjustment in health plan purchasing. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 38(3):245–259.

Kuo, R. N., Dong, Y.-H., Liu, J.-P., Chang, C.-H., Shau, W.-Y., and Lai, M.-S. (2011). Predicting healthcare utilization using a pharmacy-based metric with the whos anatomic therapeutic chemical algorithm. *Medical care*, 49(11):1031–1039.

Kutzin, J., Cashin, C., and Jakab, M. (2010). Implementing health financing reform. *World Health Organisation on behalf of the European Observatory on Health Systems and Policies*, 7(5).

Lamers, L. M. (1999a). Pharmacy costs groups: a risk-adjuster for capitation payments based on the use of prescribed drugs. *Medical Care*, 37(8):824–830.

Lamers, L. M. (1999b). Risk-adjusted capitation based on the diagnostic cost group model: an empirical evaluation with health survey information. *Health Services Research*, 33(6):1727.

Lamers, L. M. and van Vliet, R. C. (2004). The pharmacy-based cost group model: validating and adjusting the classification of medications for chronic conditions to the dutch situation. *Health policy*, 68(1):113–121.

Lamers, L. M. and Vliet, R. (2003). Health-based risk adjustment improving the pharmacy-based cost group. *The European Journal of Health Economics*, 4(2):107–114.

Lin, P. (2008). *Predictive Modeling of Health Care Expenditures for Medicare Beneficiaries with Alzheimer's Disease*. University of North Carolina at Chapel Hill.

Manning, W. G. and Mullahy, J. (2001). Estimating log models: to transform or not to transform? *Journal of health economics*, 20(4):461–494.

Mihaylova, B., Briggs, A., O'Hagan, A., and Thompson, S. G. (2011). Review of statistical methods for analysing healthcare resources and costs. *Health economics*, 20(8):897–916.

Ministerstvo zdravotnictvi CR (2007). `http://www.mzcr.cz/Legislativa/dokumenty/leciva_5619_2493_11.html`.

Ministry of Health (2012). Directive 262/2012.

Mossialos, E., Dixon, A., Figueras, J., and Kutzin, J. (2002). Funding health care: options for europe.

Pope, G. C., Ellis, R. P., Ash, A. S., Ayanian, J., Bates, D., Burstin, H., Iezzoni, L., Marcantonio, E., and Wu, B. (2000). Diagnostic cost group hierarchical condition category models for medicare risk adjustment. *Health Economics Research, Inc. Waltham, MA*.

Powers, C. A., Meyer, C. M., Roebuck, M. C., and Vaziri, B. (2005). Predictive modeling of total healthcare costs using pharmacy claims data: a comparison of alternative econometric cost modeling techniques. *Medical care*, 43(11):1065–1072.

Prinsze, F. J. and van Vliet, R. C. (2007). Health-based risk adjustment: Improving the pharmacy-based cost group model by adding diagnostic cost groups. *The Journal of Health Care Organization, Provision, and Financing*, 44(4):469–480.

Putnam, K. G., Buist, D. S., Fishman, P., Andrade, S. E., Boles, M., Chase, G. A., Goodman, M. J., Gurwitz, J. H., Platt, R., Raebel, M. A., et al. (2002). Chronic disease score as a predictor of hospitalization. *Epidemiology*, 13(3):340–346.

Regional council. `http://www.jihovychod.cz/en/regional-council/south-east-cohesion-region`.

Reinhardt, U. E. and Cheng, T.-m. (2000). The world health report 2000-health systems: improving performance. *Bulletin of the World Health Organization*, 78(8):1064–1064.

Roberts, M., Hsiao, W., Berman, P., and Reich, M. (2008). *Getting health reform right: a guide to improving performance and equity*. Oxford university press.

Smith, P. C. and Witter, S. N. (2004). Risk pooling in health care financing: the implications for health system performance.

Szalayova, A. (2012). Poistovne s chorlavejsim kmenom budu mat viac penazi. *Zdravotna politika*, (5/2012).

Van de Ven, W. P.Veen, S., van Kleef, R., van de Ven, W., and van Vliet, R. (2014). Improving the prediction model used in risk equalization: cost and diagnostic information from multiple prior years. *The European Journal of Health Economics*, pages 1–18.

Van de Ven, W. P., Beck, K., Buchner, F., Chernichovsky, D., Gardiol, L., Holly, A., Lamers, L. M., Schokkaert, E., Shmueli, A., Spycher, S., et al. (2003). Risk adjustment and risk selection on the sickness fund insurance market in five european countries. *Health policy*, 65(1):75–98.

Van de Ven, W. P., van Vliet, R. C., and Lamers, L. M. (2004). Health-adjusted premium subsidies in the netherlands. *Health Affairs*, 23(3):45–55.

Van Kleef, R. C., Van Vliet, R. C., and Van de Ven, W. P. (2013). Risk equalization in the netherlands: an empirical evaluation.

Van Vliet, R. C. (1992). Predictability of individual health care expenditures. *Journal of Risk and Insurance*, pages 443–461.

Veazie, P. J., Manning, W. G., and Kane, R. L. (2003). Improving risk adjustment for medicare capitated reimbursement using nonlinear models. *Medical Care*, 41(6):741–752.

Von Korff, M., Wagner, E. H., and Saunders, K. (1992). A chronic disease score from automated pharmacy data. *Journal of clinical epidemiology*, 45(2):197–203.

Wooldridge, J. (2012). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

Zhao, Y., Ash, A. S., Ellis, R. P., Ayanian, J. Z., Pope, G. C., Bowen, B., and Weyuker, L. (2005). Predicting pharmacy costs and other medical costs using diagnoses and drug claims. *Medical care*, 43(1):34–43.